19

DIGITAL ELECTRONICS

Interfacing digital and analogue systems

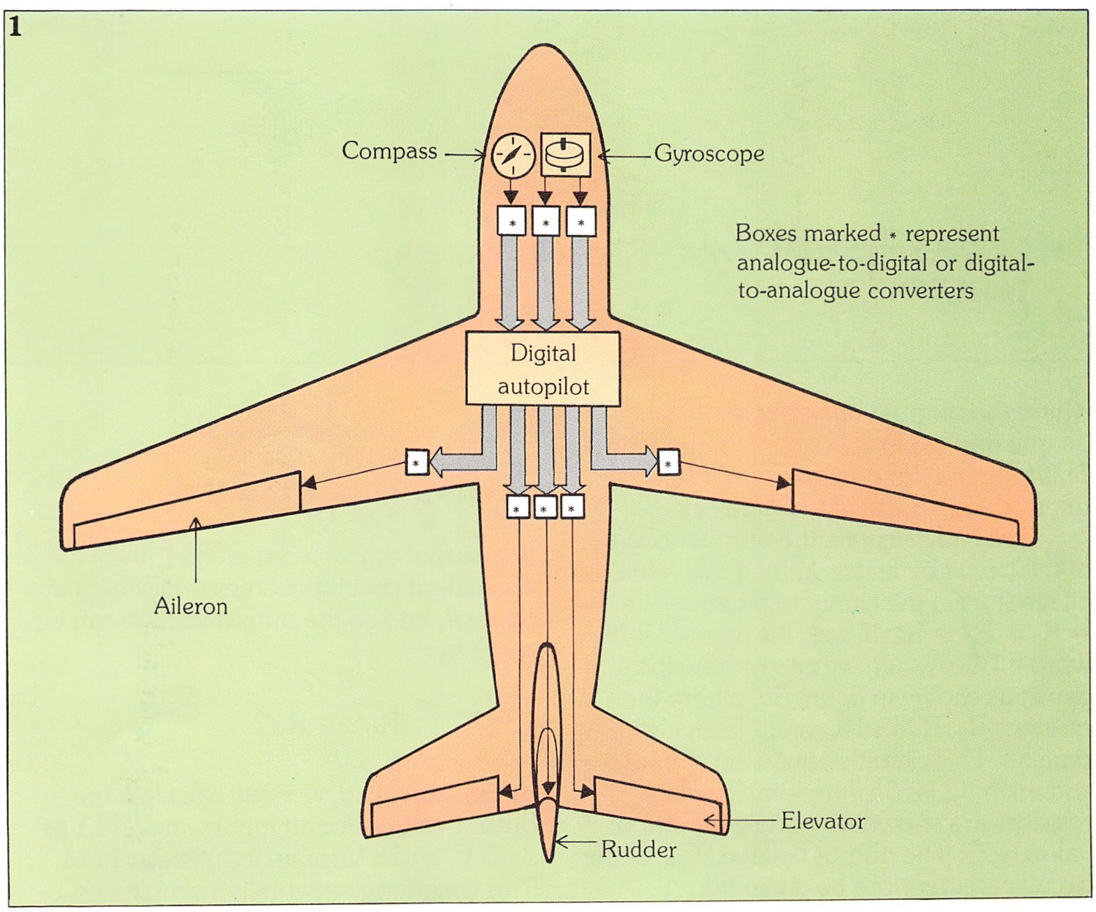
In *Digital Electronics 9* we discussed the advantages and disadvantages of digital systems, compared to analogue systems. We saw that although digital systems are inherently easier to design, display greater accuracy, are faster in operation, allow simple data storage, and can be highly integrated, they are also more complex than analogue circuits, and data transmission between parts of a system is often slower.

Because of this, it is frequently desirable to combine digital and analogue approaches to produce a system displaying the advantages of both. Such interfacing may be two-way: analogue-to-digital *and* digital-to-analogue.

An example previously seen in *Digital Electronics 9* and reproduced in *figure 1* is that of an aircraft. Here, analogue information from the compass and gyroscope is converted into digital form to be processed by the digital autopilot. The digital output from the autopilot is converted back to analogue form and used to control the rudder, ailerons and elevators which keep the aircraft on course.

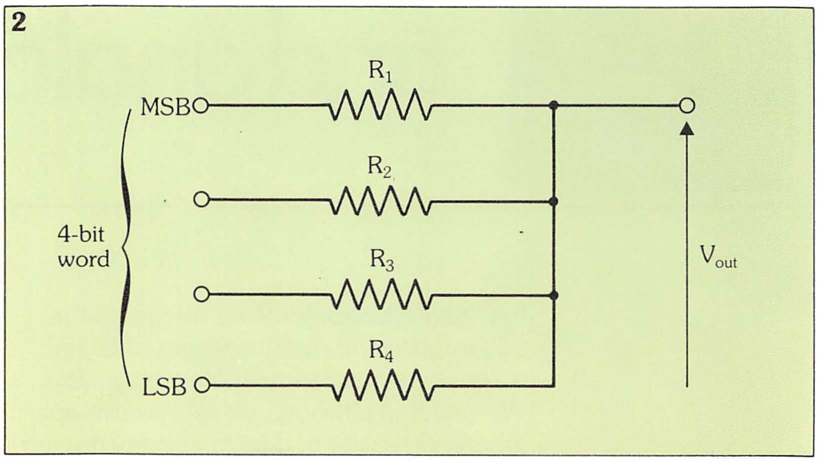
Two different categories of interfacing circuits are involved when digital and analogue systems are combined: **analogue-to-digital converters** (ADCs) and **digital-to-analogue converters** (DACs). We shall now go on to examine both of these in detail.

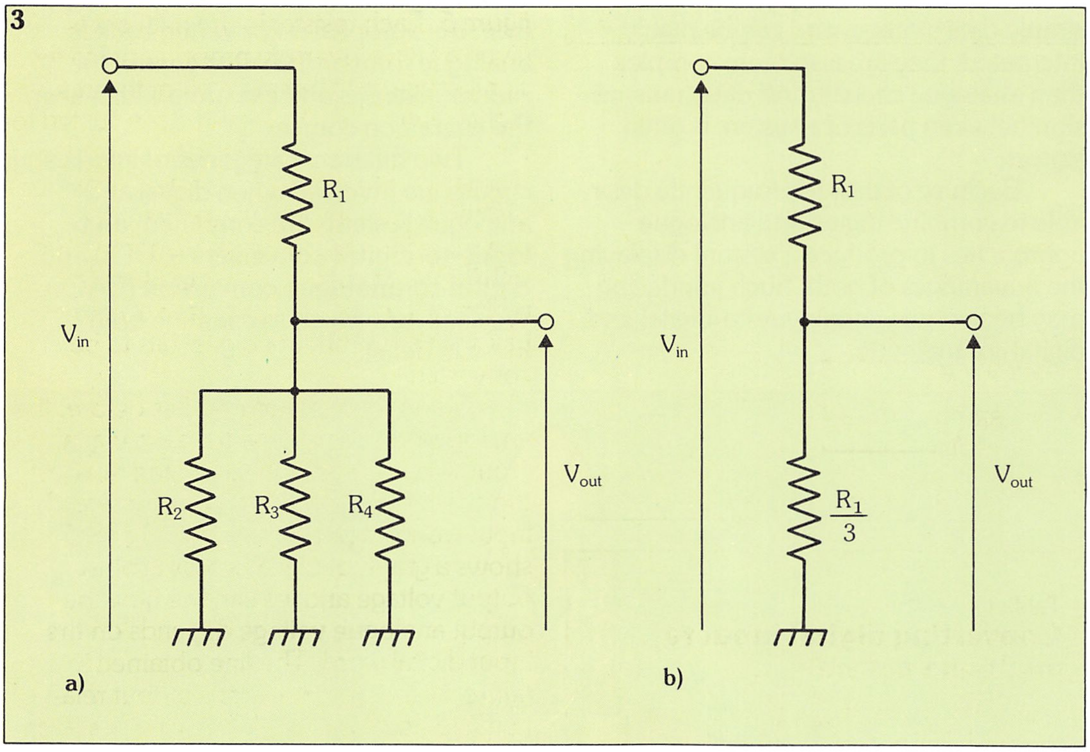
1. The automatic pilot system of an aircraft.



Digital-to-analogue conversion

The purpose of a DAC is to convert applied digitally coded words into corresponding analogue values. A simple way of doing this is shown in *figure 2*, where a 4-bit digital word is applied to four resistors, R_1, R_2, R_3 and R_4, connected at one end. The voltage level corresponding to a logic 1 bit at the input is 5 V, and logic 0 is 0 V. We can therefore see that these resistors form a type of potential divider,





2. A weighted network.

3. Calculating the output voltage of the circuit in *figure 2* if the input digital word is 0001.

where the output voltage depends not just on one applied input voltage, but on four. This type of resistor network is often referred to as a **weighted network**.

We can calculate the output voltage of the circuit by first assuming the value of all resistors in the circuit to be equal, i.e. $R_1 = R_2 = R_3 = R_4$. If, say, the applied 4-bit word is 0001, then we can redraw the circuit as shown in *figure 3a*, where the resistors R_2, R_3 and R_4 are effectively in parallel. This circuit's equivalent circuit is shown in *figure 3b* where the parallel resistors are shown as a single resistor of value $R_1/3$. The output voltage V_{out}, of the circuit can therefore be given by:

$$V_{out} = \frac{R_1/3}{R_1/3 + R_1} \times V_{in}$$

and as V_{in} is logic 1, i.e. 5 V, then:

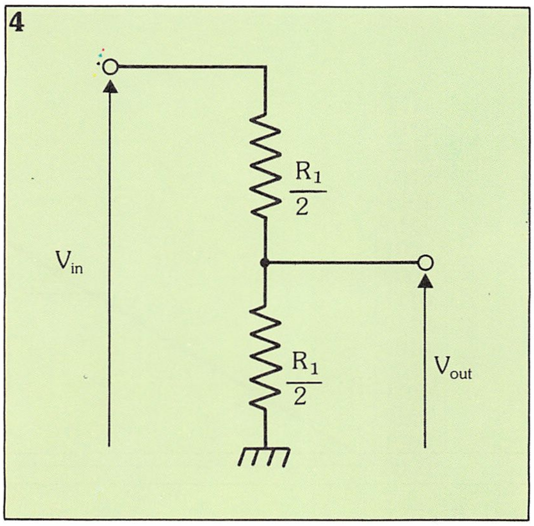
$$V_{out} = 1.25\text{ V}$$

If the input digital word is 0011, the equivalent circuit becomes that shown in *figure 4*, and so the output voltage can be given by:

$$V_{out} = \frac{R_1/2}{R_1/2 + R_1/2} \times V_{in}$$
$$= 2.5\text{ V}$$

In the same way, we may calculate the output voltage for an input word 0111 as 3.75 V, and 5 V for an input word 1111. This weighting network is a simple DAC.

4. Equivalent circuit for *figure 3a* if the input digital word is 0011.



5. A binary weighted network.

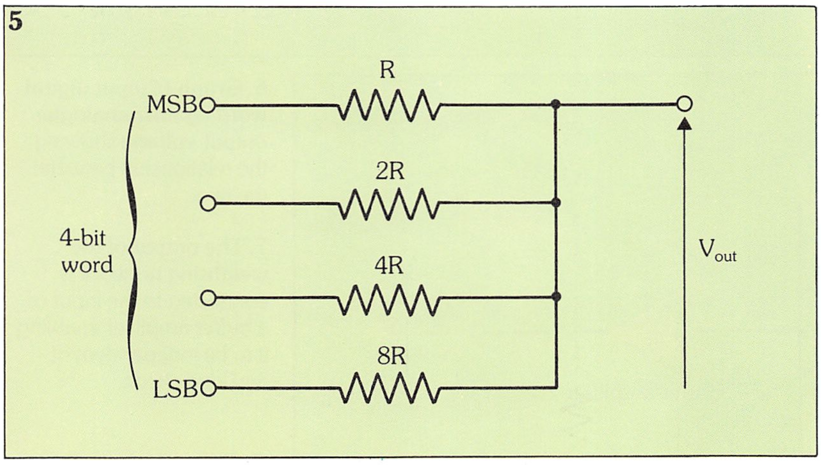


Table 1
Converting digital input to analogue output

Digital word	Output voltage, V_{out}
0000	0 = 0 V
0001	$\frac{1}{15} \times V_{in}$ = 0.33 V
0010	$\frac{2}{15} \times V_{in}$ = 0.67 V
0011	$\frac{1}{5} \times V_{in}$ = 1 V
0100	$\frac{4}{15} \times V_{in}$ = 1.33 V
0101	$\frac{1}{3} \times V_{in}$ = 1.67 V
0110	$\frac{2}{5} \times V_{in}$ = 2 V
0111	$\frac{7}{15} \times V_{in}$ = 2.33 V
1000	$\frac{8}{15} \times V_{in}$ = 2.67 V
1001	$\frac{3}{5} \times V_{in}$ = 3 V
1010	$\frac{2}{3} \times V_{in}$ = 3.33 V
1011	$\frac{11}{15} \times V_{in}$ = 3.67 V
1100	$\frac{4}{5} \times V_{in}$ = 4 V
1101	$\frac{13}{15} \times V_{in}$ = 4.33 V
1110	$\frac{14}{15} \times V_{in}$ = 4.67 V
1111	V_{in} = 5 V

Binary weighted network

It would appear from this description that we have a DAC which converts an input 4-bit digital word into a corresponding analogue signal, the value of which depends on the value of the digital word. However, the analogue output voltage is not uniquely related to the input word; for example the word 1001 produces the same output voltage as the word 0011, and the word 1110 produces the same output voltage as the word 0111, even though the digital words have different meanings.

The problem is overcome by using a **binary weighted network** as shown in *figure 5*. Each resistor in the network is given a value inversely proportional to the significance of the applied bit of the input digital word. Thus the resistor connected to the most significant bit (MSB) of the data word may be of value R, the next resistor is 2R, the next is 4R and the last resistor, connected to the least significant bit (LSB), is 8R. The actual value of R in ohms is arbitrary and should be chosen on the basis of the number of digital bits to be converted.

Using the same method as before, the analogue output voltage for each digital input word can be calculated; the result of the conversions for all 16 possible digital input words is given in *table 1*. *Figure 6* shows a graph of digital words against output voltage and we can see how the output analogue voltage depends on the input digital word. The line obtained in *figure 6* is not the true input/output relationship however – we'll see why not later – but is adaquate for now.

Usually, the output of the weighting network is connected to the input of a buffer amplifier as shown in *figure 7*. In this way, the circuit's output voltage becomes independent of any load circuit which may otherwise affect it.

This DAC circuit may be used to convert digital words of any number of bits, by adding a further resistor for every extra bit according to the binary weighting. The disadvantage is that for every extra bit, the range of required resistors doubles. Consider a 13-bit DAC: if we assume a value of 10 kΩ for the resistor corresponding to the MSB, then the LSB resistor value must be 40.96 MΩ. Resistors of that sort of

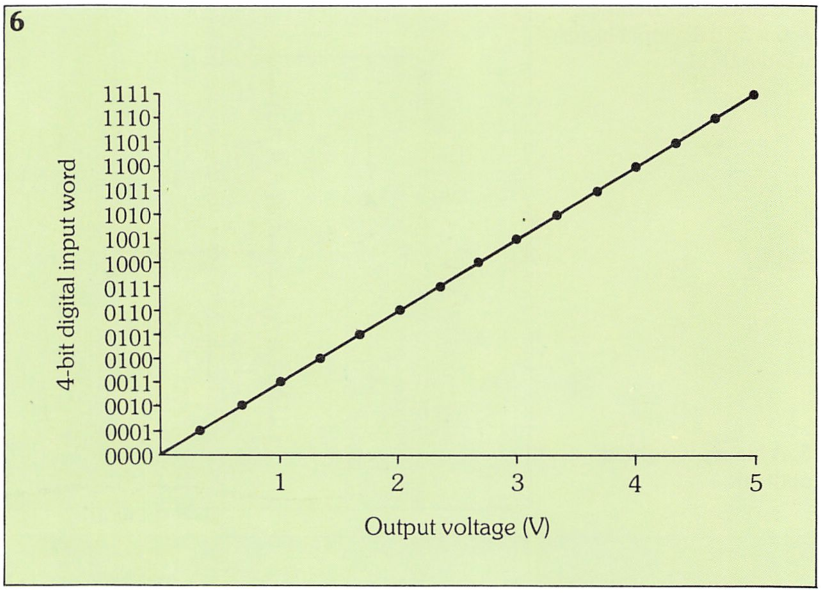
range are difficult to manufacture accurately, so another method of converting digital signals to analogue must be found for digital words with large numbers of bits.

R-2R ladder DACs

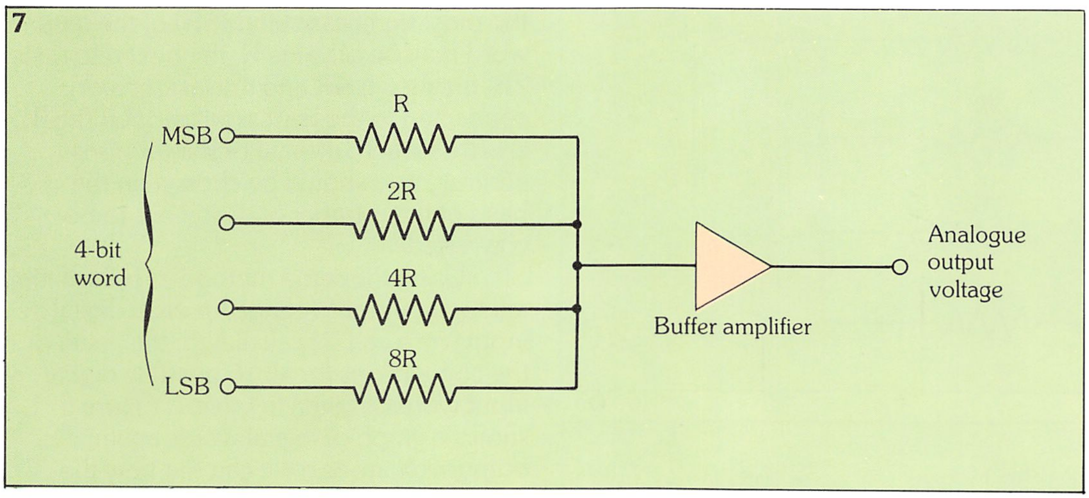
A second technique often used in DACs is shown in *figure 8a*, the **R-2R ladder network**. Also shown is a buffer amplifier of the same type as that in *figure 7*.

The circuit works on the principle that the pair of 2R resistors at the end of the ladder can be considered in parallel, forming an overall resistor of value R. Thus, the circuit can be redrawn as in *figure 8b*. The two resistors of value R are in series producing a total of 2R which, when considered in parallel with the 2R resistor, gives a resistor of R. This procedure carries on down the ladder until the circuit can be considered as in *figure 8c*.

For an equal number of digital input bits, this circuit uses twice the number of resistors compared to the binary weighted network, but the use of only two values of resistors means that the circuit can handle any number of input bits, simply by extending the resistor ladder.



6. Graph of input digital word against analogue output voltage showing the relationship between the two.



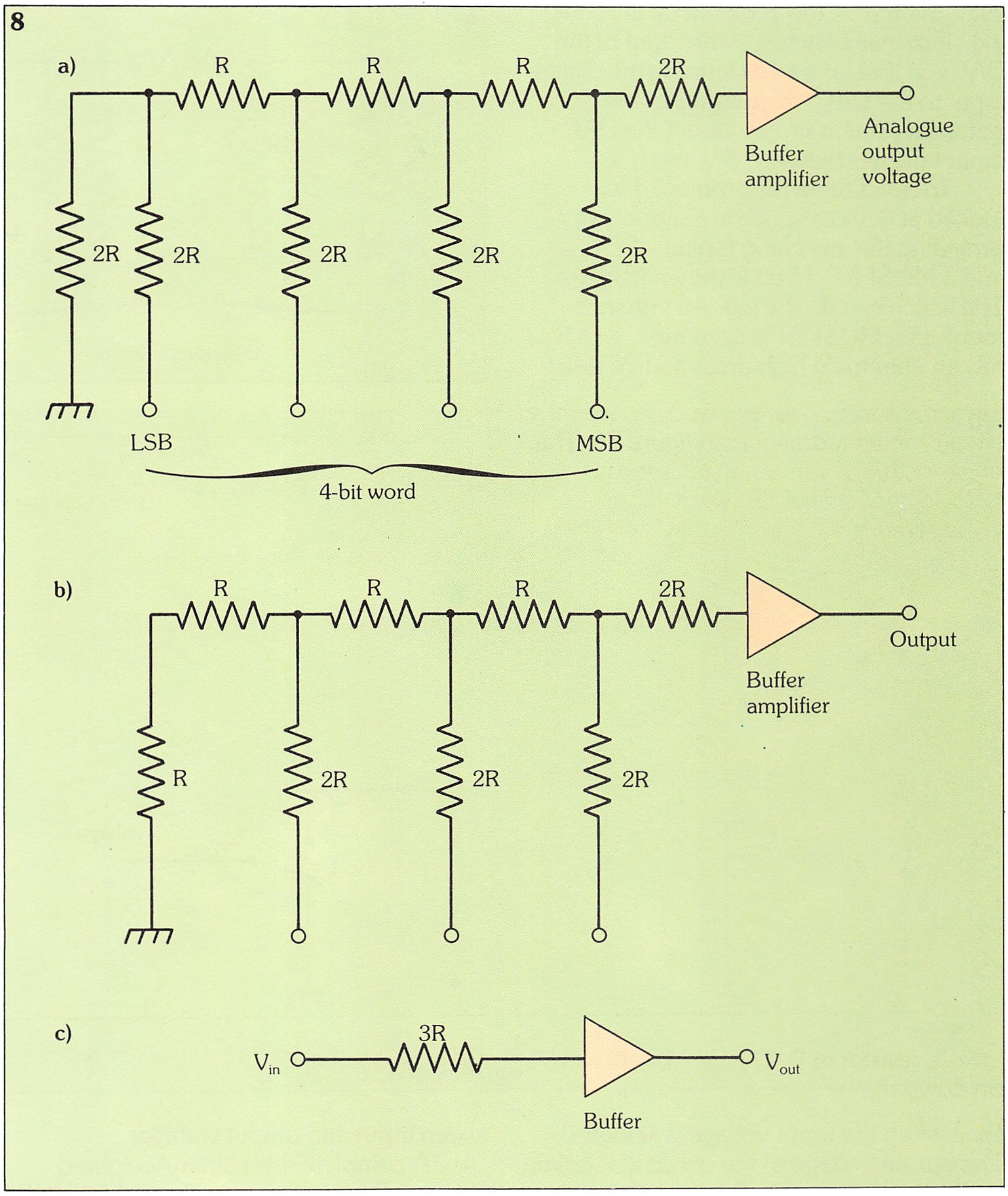
7. The output of the weighting network is connected to the input of a buffer amplifier enabling it to be independent of any load circuit.

Below: a single engineer controls an entire System X Exchange at Arrington in Suffolk.

Combining the binary weighted network and the R-2R ladder

A compromise solution between the binary weighted and R-2R ladder DACs is often used and is shown in *figure 9*. Here, we have an 8-bit DAC and the inputs are arranged in two groups of four. Digital-to-analogue conversion of each group is undertaken by the two binary weighted

8. Digital-to-analogue conversion using an R-2R ladder network.



networks and a link is made between the two groups by an attenuation resistor, R_A. The value of this resistor is chosen to binary weight the left group by reducing the current produced by a factor of 16. This would occur when resistor $R_A = 8R$.

The circuit of *figure 9* is also useful when the digital input words are not of natural binary format but, say, of binary coded decimal format. All that is required is to change the value of resistor R_A, reducing the attenuation factor from 16 to 10. In this case the value of resistor R_A is 4.8R.

Switching inputs

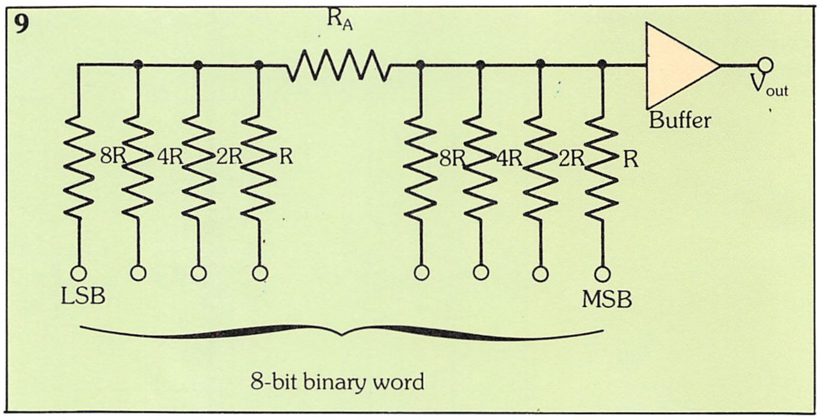
The DACs in *figures 2, 8* and *9* could not be used as they stand for converting digital signals to analogue signals because the circuits providing the digital input words do not produce accurately defined output voltages. If you remember, the outputs of logic gates, for example, are defined only within *limits* of voltages (e.g. above 3.5 V corresponds to logic 1, below 2 V corresponds to logic 0).

The DACs we have seen so far require exact input voltages (i.e. 5 V and 0 V) for correct operation. We can, how-

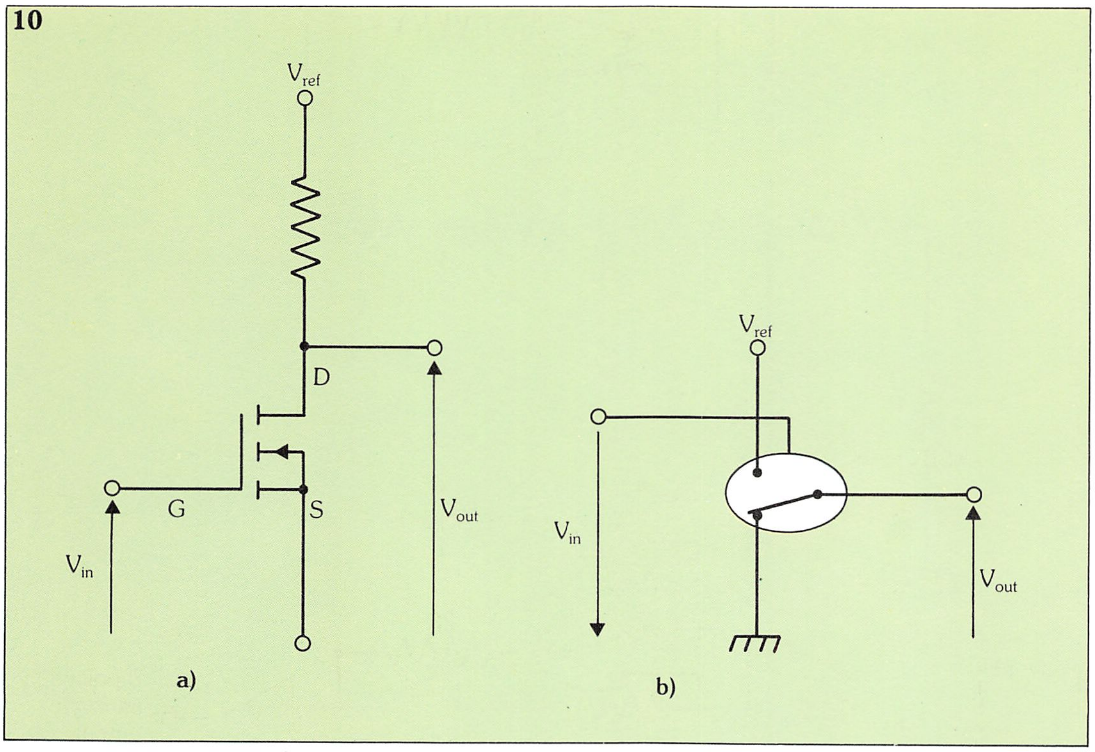
ever, overcome this problem by introducing electronic switches at the input of the DACs so that instead of forming the digital input to the DAC, the outputs of the preceding circuit simply *switch* the DAC input voltages between 5 V and 0 V.

In *Solid State Electronics 14* we looked at devices which are capable of providing this switching facility – FETs – and a MOSFET circuit is shown in *figure 10a* which can do the job. An enhancement-type MOSFET is used here, which has an extremely high drain-source resistance when the input voltage is at logic 0. The output voltage of the circuit at this time is effectively the reference voltage (minus a negligible voltage drop across the resistor, R_1). However, when a logic 1 input is applied, the MOSFET switches from an extremely high resistance to a low resistance and the output voltage becomes 0 V (plus a negligible voltage drop across the transistor).

Figure 10b shows the symbol we will use to represent this electronic switch. It is simply a double-throw (i.e two-way) switch, switching between the reference voltage and ground, controlled by an applied input voltage. The different arrowhead directions indicate the inversion between input and output voltages.



9. **Combining the binary weighted** network and the R-2R ladder.



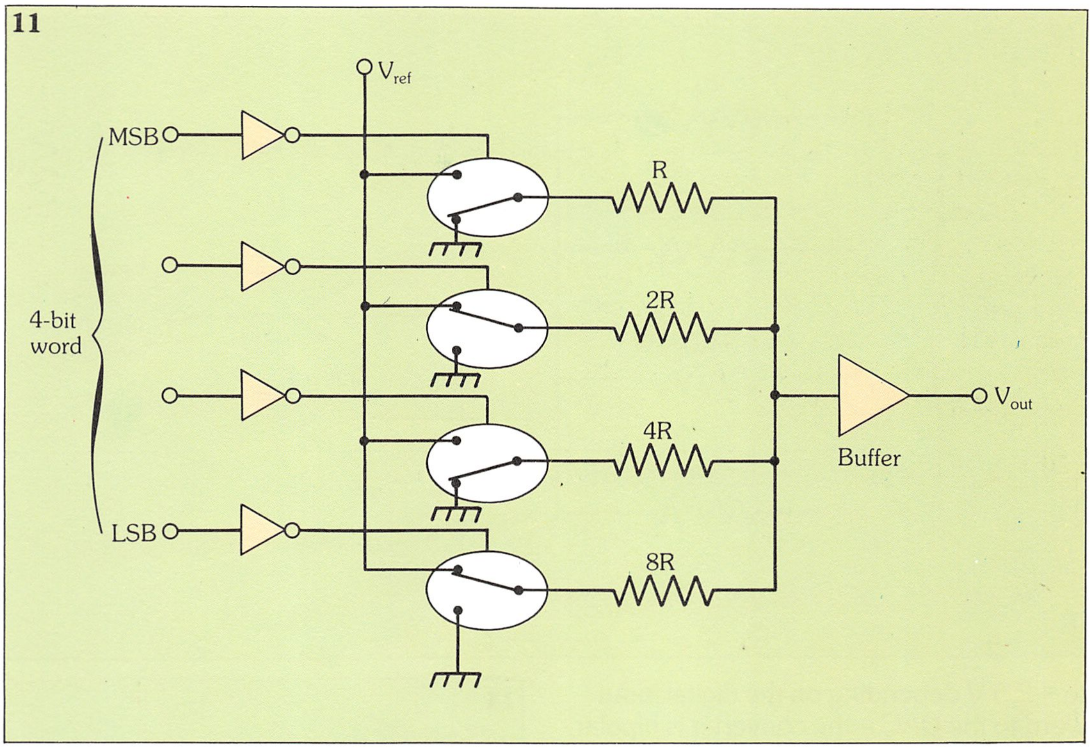
10. **(a) An enhancement-type MOSFET** circuit used to switch the DAC input voltages between 5 V and 0 V; **(b)** circuit symbol for the switch.

A complete 4-bit binary weighted network DAC is shown in *figure 11* which features four MOSFET electronic switches with a single reference voltage source. Also included are the four required inverters to re-invert the applied bits of the digital words.

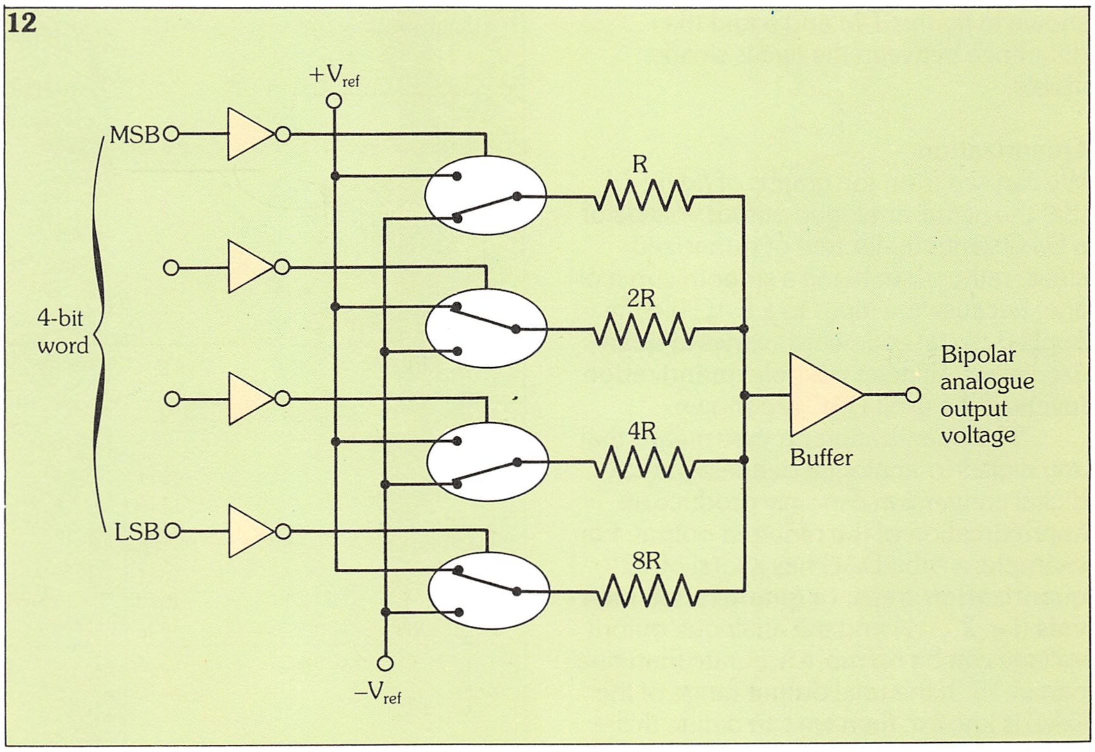
Unipolar and bipolar conversion

The analogue output voltages of all the DACs we have seen so far have been of a single polarity (i.e. positive), and because of this are termed **unipolar**. *Figure 12* shows a DAC which will produce an output of negative *and* positive polarity, that is, a **bipolar** DAC. Such converters are

11. A 4-bit binary weighted network DAC featuring four MOSFET electronic switches with a single reference voltage source.

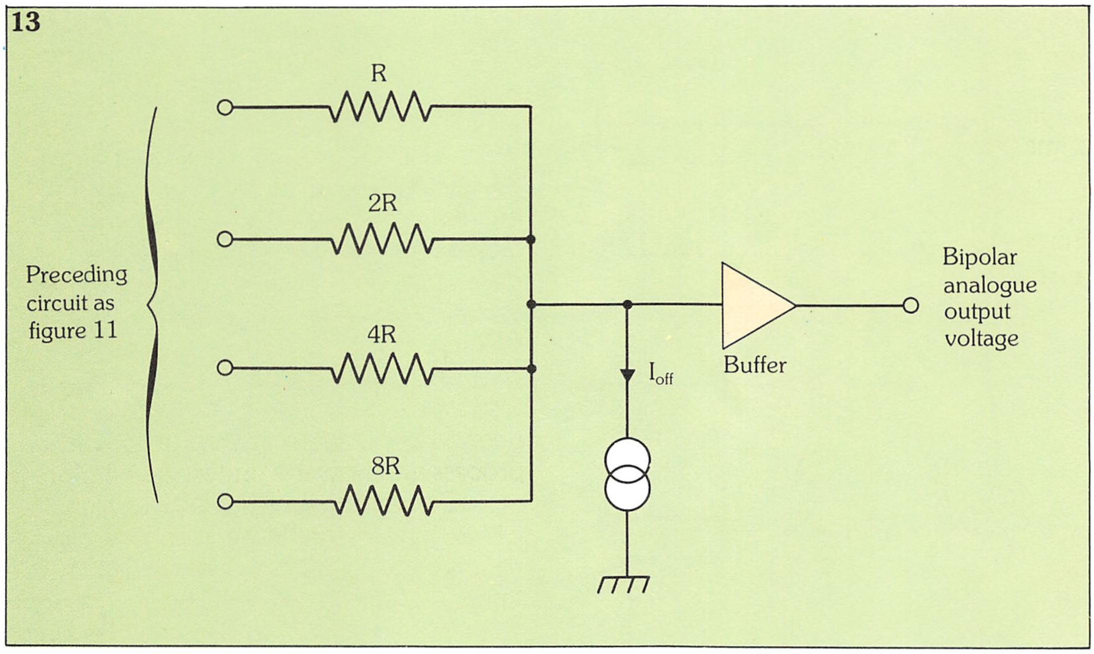


12. A bipolar DAC.

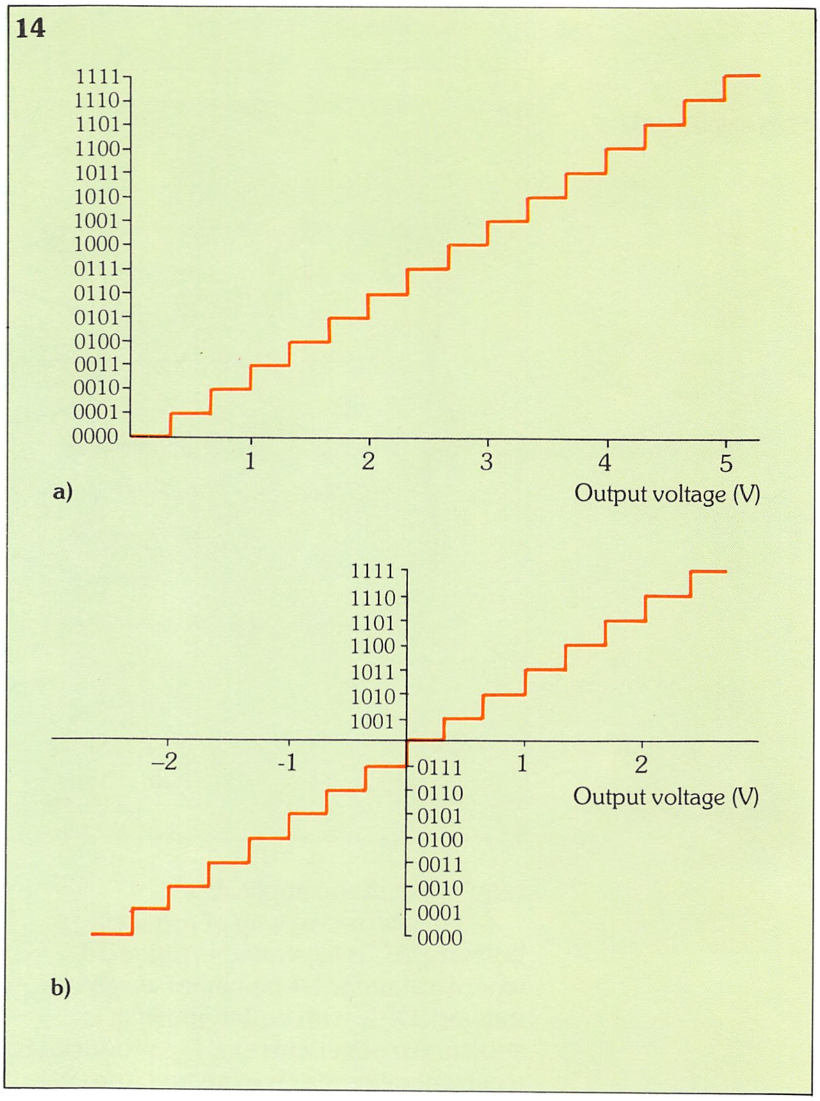


required in many applications.

An alternative way of producing a bipolar DAC is illustrated in *figure 13* where a standard 4-bit binary weighted network DAC with buffer amplifier is shown. An **offset current**, I_{off}, produced by a current generator is extracted from the amplifier input circuit. The offset current produces a fixed output voltage which acts as a voltage reference to which all voltages created by the digital inputs after conversion are added. So, if the offset current produces a fixed output voltage of −2.5 V, the output voltage may vary from −2.5 V



13. An alternative way of producing a bipolar DAC.



14. Digital input *vs* analogue output voltage for **(a)** unipolar; and **(b)** bipolar DACs.

to +2.5 V depending on the digital input word to the DAC – the converter is bipolar. Unipolar and bipolar DAC outputs are shown in *figures 14a* and *b* and the difference between the two is clearly visible.

Quantization

We can see from the graphs of *figure 14* that the actual analogue output voltage of a DAC varies in discrete or **quantized** steps, rather than being a smooth curve or line, because the *input* to a DAC (i.e the applied digital codeword) varies in quantized steps. Sixteen possible **quantization levels** of the 4-bit DAC are shown.

The quantization process means that any digital-to-analogue or analogue-to-digital conversion can only produce an approximation of the required output. For example, a 4-bit DAC has a total of 15 **quantization steps**, or **quantization intervals** (i.e. $2^4 - 1$) and the analogue output voltage can be no more accurate than one part in 15. If the total output range of the DAC is known, then we can define this inaccuracy as a **quantization error**, where the error is plus or minus half the quantization interval. So, for a total voltage range of 5 V, a 4-bit DAC has a quantization error of:

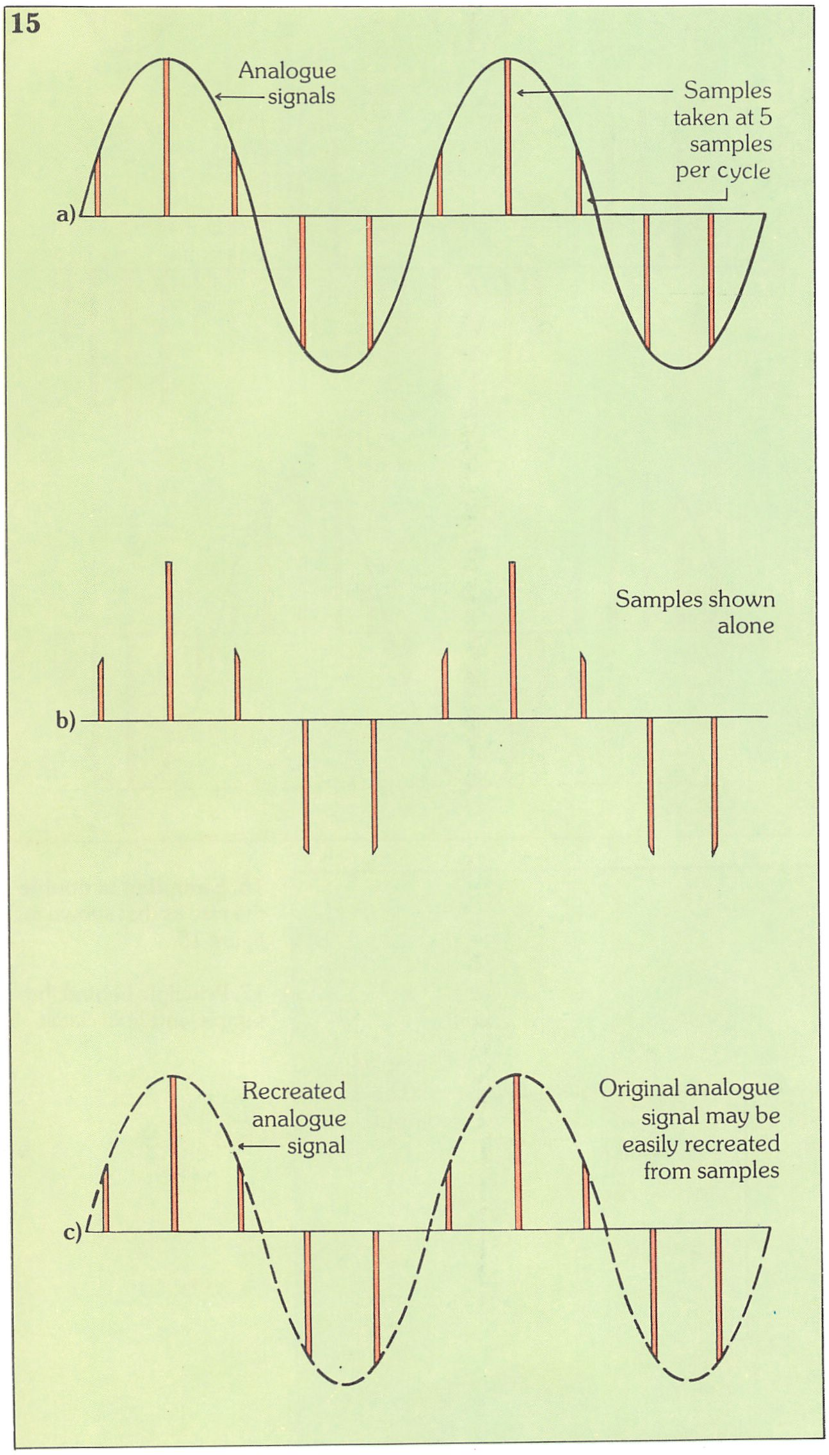
$$\frac{5}{2 \times 15} = \pm 0.17\,\text{V}$$

On the other hand, an 8-bit DAC with a total output voltage range of 5 V has a quantization error of:

$$\frac{5}{2 \times (2^8-1)} = \frac{5}{2 \times 255} = \pm 0.01 \text{ V}$$

This shows that the greater the number of bits converted, the smaller the quantization error.

15. Sampling.



Analogue-to-digital conversion

We have seen that the conversion of digitally coded signals to a corresponding analogue signal is a fairly simple task. Conversion of an analogue signal to a number of digital bits is, however, more complex and many of the methods used in ADCs are based on DACs. We shall examine the basic circuits and principles of ADCs in *Digital Electronics 20*, but we need to know a little about the conversion process and how it is undertaken, before this.

Converting an AC analogue signal which, by its very nature, is continually changing in value, to a discrete digital signal requires that the analogue signal's value must first be measured regularly, and the value at each measurement then converted to the corresponding digital form. The process of measuring values at regular intervals is known as **sampling** and the rate at which samples are taken is the **sampling rate**.

Figure 15a shows an arbitrary analogue (a sine wave) which is sampled at a rate of 5 samples per cycle. In *figure 15b* we can see the individual samples, the values of which can be translated to a digital signal. These sample values may be used to recreate approximately the original sine wave, shown by the broken line in *figure 15c*.

Obviously, the faster the sampling rate the closer the sampled values represent the complete analogue signal. *Figure 16a*, for example, shows the same sine wave as before, sampled 10 times per cycle; *figure 16b* shows the samples; and *figure 16c* illustrates how the sine wave can be recreated.

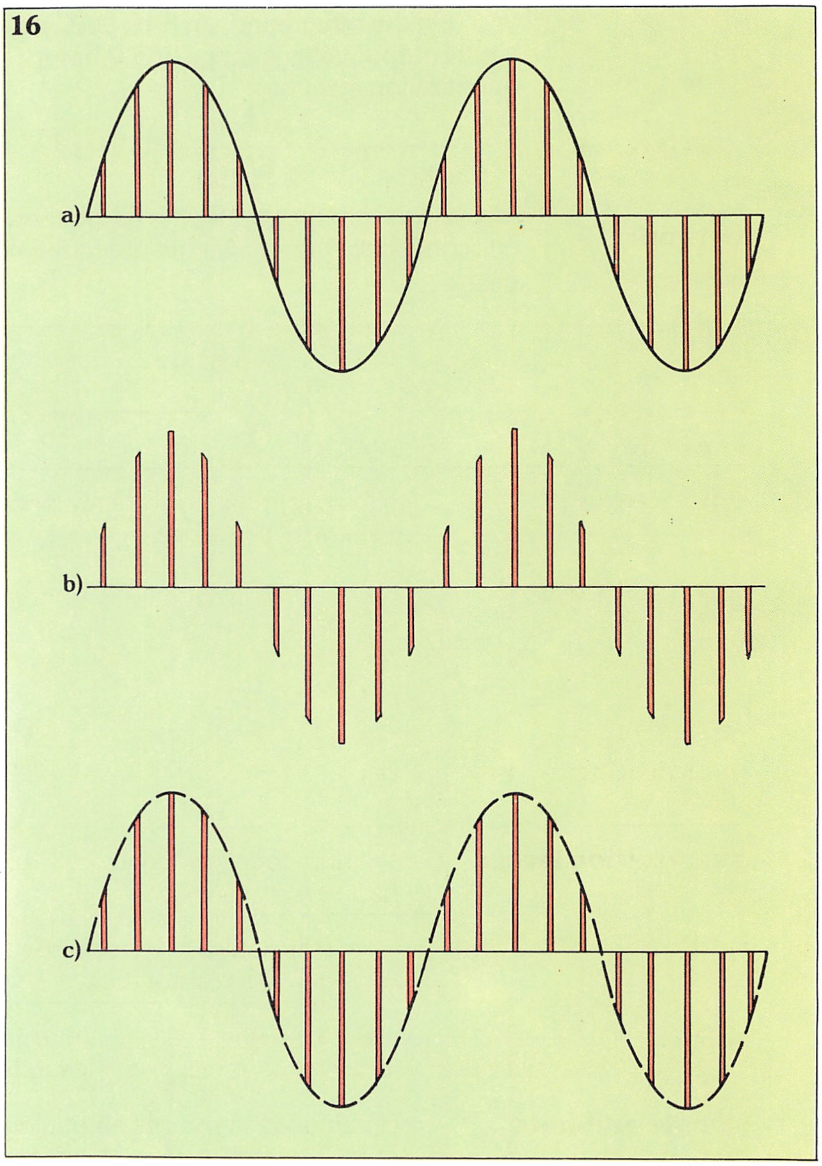
However, sampling the analogue signal as fast as possible is sometimes inefficient and unnecessary, and it is possible to recreate an analogue signal using far fewer samples than in these two examples. The minimum sampling rate is related to the actual analogue signal by the **sampling rule** which states that: any signal with a bandwidth from DC to f (Hz) can be completely recreated from regular samples taken at a rate of at least 2 f samples per second. So, a sine wave of, say, 1 kHz as

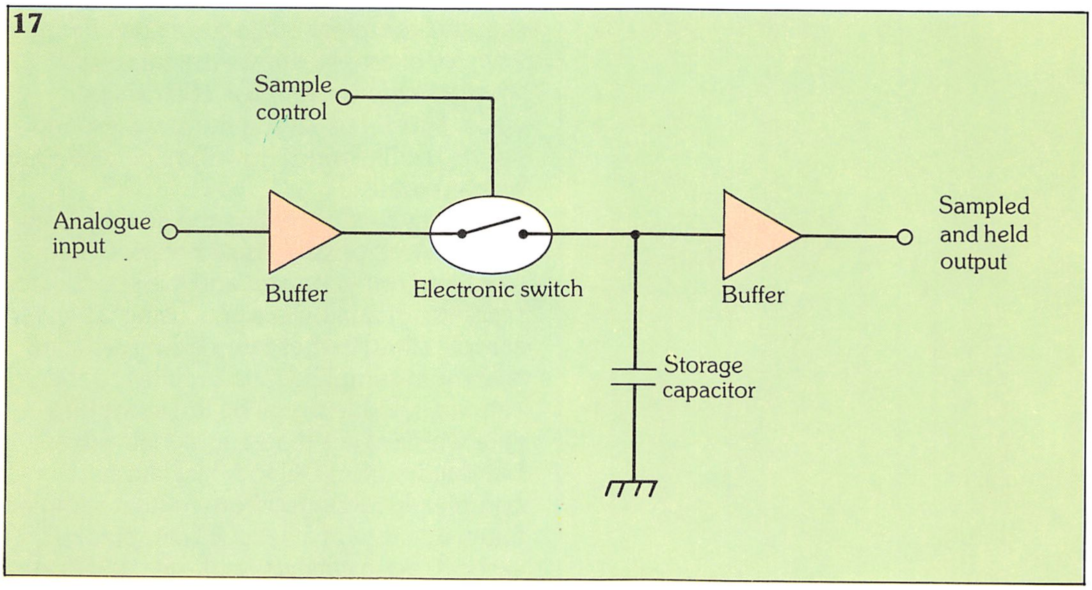
shown in *figure 16* must be sampled at a rate of at least 2000 samples per second in order that it may be recreated successfully from the sample values. Similarly, an analogue signal corresponding to music, and consisting of a collection of frequencies up to 15 kHz, may be represented by and recreated from a set of samples taken at 30,000 samples per second.

Sample and hold

Conversion from the analogue sample voltage to the digital output takes a finite time. During this time the analogue signal may vary, affecting the conversion process. In many practical ADCs therefore, the analogue sample voltage is stored in some way (generally across a capacitor) until the conversion is complete. The principle of such a **sample-and-hold circuit** is shown in *figure 17*, where an electronic switch is used to connect the sampled analogue signal to the capacitor. When the capacitor voltage equals the analogue voltage, the switch disconnects the capacitor which holds its stored charge. The output voltage of the circuit thus equals the sampled voltage and remains so, ideally until the next sample, whereupon the switch once again connects the capacitor to the input sampled voltage.

Buffer amplifiers are included at the circuit's input and output to prevent loading of the analogue signal source by the capacitor, and also loading of the capacitor by any following circuit.





16. Sampling at double the rate as that shown in *figure 15*.

17. Principle behind the sample-and-hold circuit.

Glossary

analogue-to-digital converter (ADC)	a circuit which converts a continually varying analogue signal into a number of digital bits
binary weighted network	a resistive network used as the basis of a type of DAC, in which the value of each resistor in the network is inversely proportional to the significance of each bit in the applied digital signal
bipolar DAC	a digital-to-analogue converter, the output of which can be a positive or negative analogue voltage depending on the digital input
digital-to-analogue converter (DAC)	a circuit which converts applied digital signals into a varying analogue signal
offset current	current applied to, or extracted from, the input of an amplifier which generates a fixed voltage at the amplifier output
quantization error	error which the quantization process causes in digital-to-analogue and analogue-to-digital conversion. It results from the fact that the digital signal, corresponding to an analogue signal value, is only an approximation
quantization interval, quantization steps	the number of intervals between quantization levels of a possible range of digital signals. A 4-bit range has a maximum total of 15 (i.e. 2^4-1) quantization intervals
quantization levels	the number of levels of a possible range of digital signals. A 4-bit range has a total of 16 (i.e. 2^4) quantization levels
R-2R ladder network	a resistor network, used as the basis of a type of DAC, in which only two values of resistor are used, regardless of the number of applied bits
sampling	process of regularly measuring an analogue signal to find its instantaneous values at the regular sample times
sample-and-hold circuit	circuit used to sample an analogue signal, and hold or store the value of each sample to allow time for the ADC to function
sampling rate	the rate at which samples of an analogue signal are taken
sampling rule	rule which gives the minimum sampling rate of an analogue signal to allow later recreation of the signal
unipolar converter	a DAC which gives an output of single polarity, i.e. positive or negative, but not both

BASIC THEORY REFRESHER

ELECTRICAL TECHNOLOGY

AC in inductors and capacitors

Having seen the way in which AC behaves in circuits made up of resistances, we must now consider its effect on inductors and capacitors. For the sake of simplicity, we shall look at idealised forms of these two elements first of all, and later find out how the real circuit components differ from the idealised models.

AC in an inductor

In an earlier *Basic Theory Refresher*, we found that the voltage, v, needed to drive a current, i, through an inductor is proportional to the rate of change of the current with time (*figure 1*). Mathematically, if the current increases from i_1 to i_2 in a very short time interval t, then:

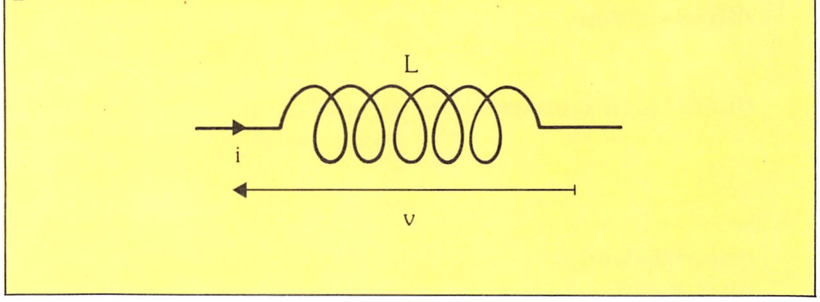
$$v = L\left(\frac{i_2 - i_1}{t}\right)$$

where L is the self inductance of the inductor.

Figure 2a shows a sinusoidal current, i (represented by the black curve), where $i = \hat{I} \sin \omega t$. We can take a number of very small intervals along the time axis and measure the difference in current at the start and end of each time interval. By multiplying this value by L, we can determine the value of the voltage needed to keep this current flowing; this is shown in red in *figure 2a*. You'll notice that the voltage is at its maximum value when the current is zero; positive when the current is increasing; and negative when the current is decreasing. Again, when the current is maximum, the required voltage is zero.

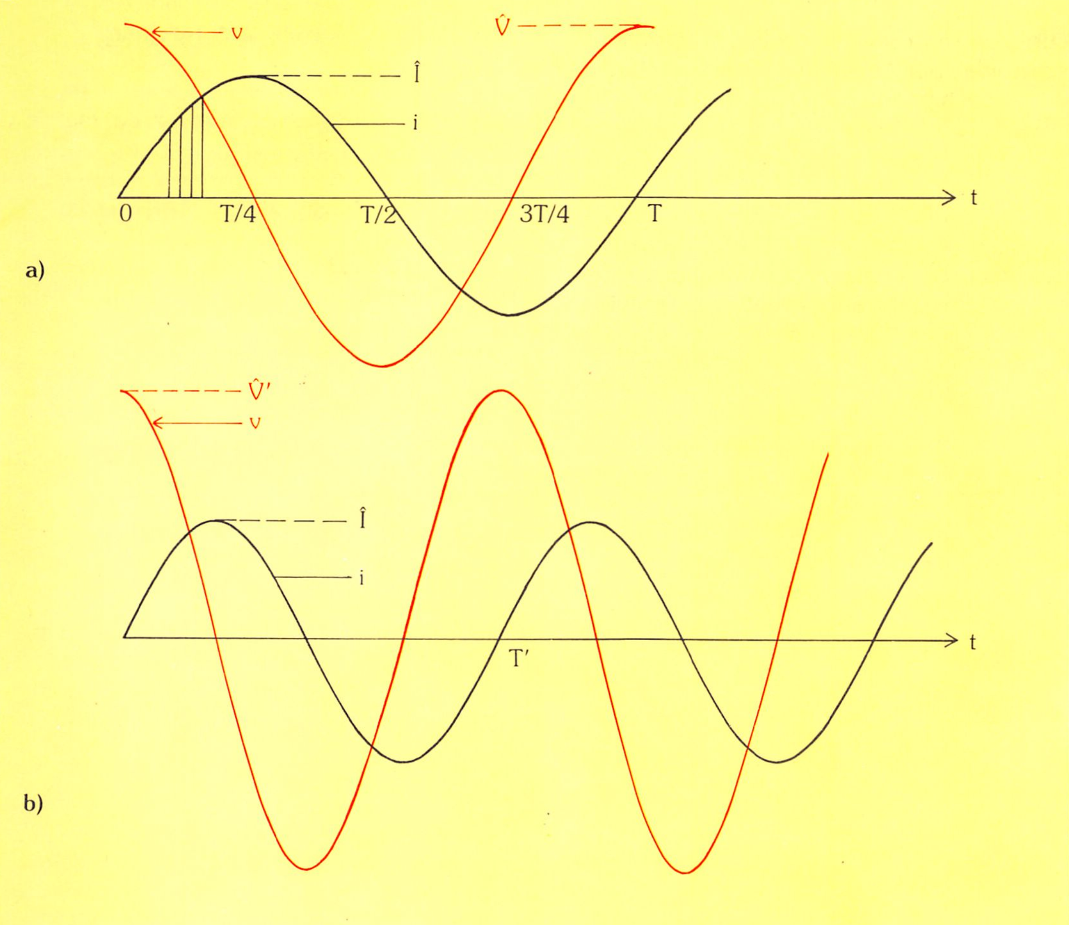
We can see from *figure 2*, that when the

1

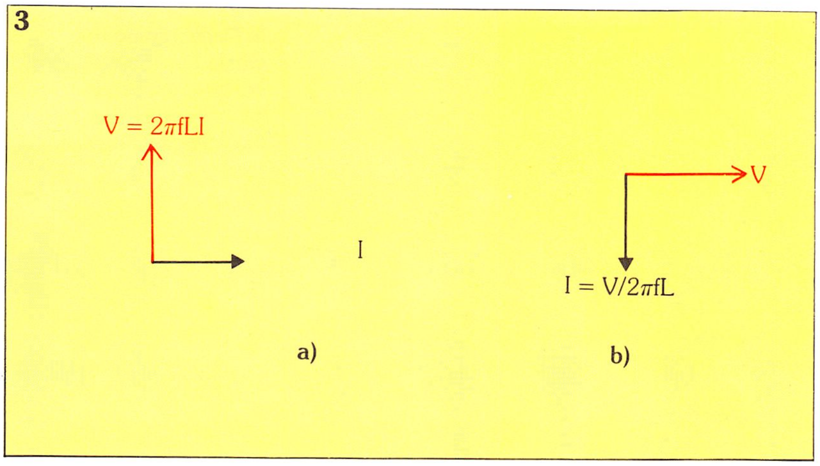


1. The voltage required to drive a current through an inductor is proportional to the rate of change of current with time.

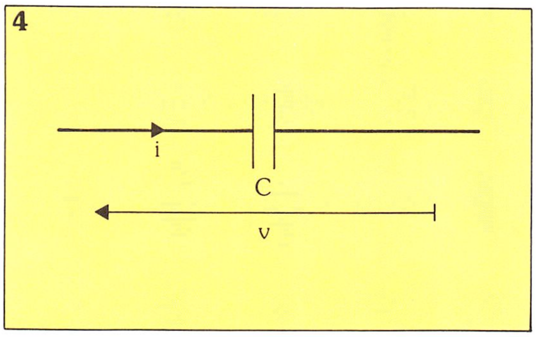
2



2. The time axis is divided into very small intervals. The current differences at those intervals are multiplied by the inductance to give the voltage.



3. Phasor diagrams for an inductor.



4. A capacitor, C, with a sinusoidal voltage, V, applied across it.

current is sinusoidal, the voltage is also sinusoidal, but displaced by a time T/4 – a quarter of a period of cycle. We can therefore write this as:

$$\begin{aligned} v &= \hat{V}\cos\omega t \\ &= \hat{V}\sin(\omega t + \pi/2) \end{aligned}$$

where $\hat{V}$ is the maximum value of the voltage.

If we now increase the frequency to a new value ω', keeping the magnitude of the current $\hat{I}$ the same and repeating the previous procedure, we obtain the graph in *figure 2b*. We can see that when the current is passing through zero, its rate of change is faster than in the first case, so the magnitude of $\hat{V}'$ will therefore be greater than was previously required.

Reactance of an inductor

We can show that the magnitude of the required voltage at any frequency is proportional to the frequency as well as to the inductance. Thus:

$$\begin{aligned} \hat{V} &= \omega L\hat{I} \\ &= 2\pi f L\hat{I} \end{aligned}$$

If we use the rms value of voltage V, and current I, we can say that:

$$V = 2\pi f L I$$

since:

$$V = \frac{\hat{V}}{\sqrt{2}}$$

and:

$$I = \frac{\hat{I}}{\sqrt{2}}$$

So, we can now write this equation in the form:

$$V = XI$$

This gives us an expression reminiscent of Ohm's law, but we must remember that V and I are out of phase, with I lagging V.

The quantity X in the equation above is termed the **reactance** of the inductor:

$$X = 2\pi f L$$

This is the ratio of the rms value of a sinusoidal voltage and the resulting sinusoidal current – which are $\pi/2$ out of phase with each other. Reactance, like resistance, is measured in ohms. We can see that the reactance of an inductor increases as the frequency increases.

Phasor diagram for an inductor

If we take the current as our reference direction, then we can draw the phasor diagram shown in *figure 3a*. Here, I represents the current and $V = 2\pi f LI$ is the applied voltage leading it by 90° (i.e. $\pi/2$). On the other hand, taking the voltage as the reference direction will give us the phasor diagram shown in *figure 3b*. As you can see, these two diagrams are identical, as the choice of reference direction is entirely arbitrary.

Example

Let's find the rms value of the current that flows through an inductor of 2 mH, which is connected to a sinusoidal voltage of rms value 5 V, at a frequency 1 kHz. The reactance:

$$\begin{aligned} X &= 2 \times \pi \times 1 \times 10^3 \times 2 \times 10^{-3} \\ &= 12.57\Omega \end{aligned}$$

so the current:

$$\begin{aligned} I &= \frac{5}{12.57} \\ &= 0.398\text{ A} \\ &= 398\text{ mA} \end{aligned}$$

If we increase the frequency to 1 MHz, we find that the reactance is now:

$$\begin{aligned} X &= 2 \times \pi \times 1 \times 10^6 \times 2 \times 10^{-3} \\ &= 12.57\text{ k}\Omega \end{aligned}$$

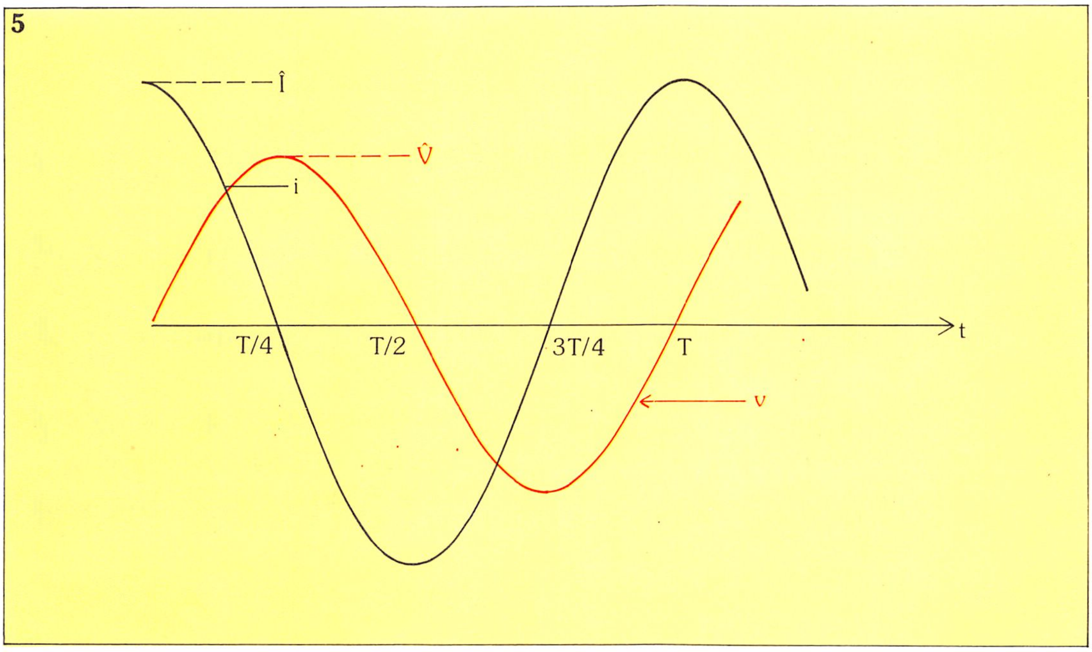
and:

$$I = 398\,\mu\text{A}$$

As the frequency increases 1000 times, so the reactance also increases 1000 times and the current falls by a factor of 1000. Remember however, that current lags voltage in both cases.

Alternating voltage across a capacitor

Let's look at a capacitor of value, C, that has a sinusoidal voltage $v = \hat{V}\sin\omega t$ applied across it, as shown in *figure 4*. We have seen that the charge, q, on a capacitor is related to the



5. Determining the current by dividing the time axis into small intervals and measuring the voltage difference.

voltage by:

$$q = Cv$$

And as the current, i, flowing in a circuit is the rate of change of charge, we may state that the current flowing into a capacitor is C times the rate of change of voltage.

We can draw a wave diagram to show this, like that in *figure 5*. We start with the voltage, v, and derive the current by taking small increments of time like we did for the inductor. This gives the waveform shown, where:

$$i = \hat{I} \sin(\omega t + \pi/2)$$

Here we see that the current leads the voltage by $\pi/2$. By a similar argument to that used for the inductor:

$$\hat{I} = \omega C\hat{V}$$
$$= 2\pi fC\hat{V}$$

which using rms values gives us:

$$I = 2\pi fCV$$

Reactance of a capacitor

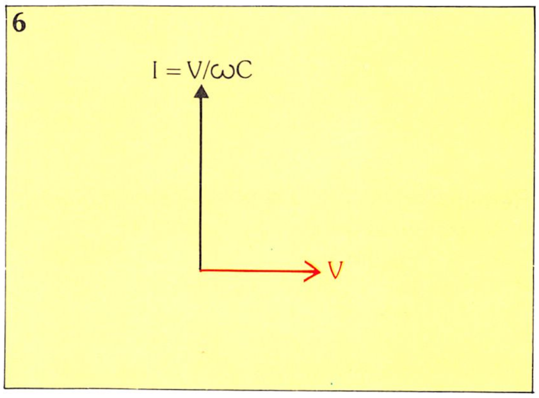
As for the inductor, we can write the equation for a capacitor's reactance in a form similar to Ohm's law:

$$V = XI$$

where:

$$X = \frac{1}{2\pi fCV}$$

is the reactance, defined as the ratio of a voltage to a current, 90° out of phase. Note that the reactance of a capacitor is inversely proportional to frequency, and that the current leads the voltage by 90°.



6. Phasor diagram for a capacitor.

Table 1
Properties of capacitors and inductors

Property \ Circuit	Inductance	Capacitance
Phase relationship	Current lags voltage by 90°	Current leads voltage by 90°
Variation of reactance	Increases with frequency	Decreases with frequency
Reactance at low frequencies	Very small (approximates a short circuit)	Very large (approximates a open circuit)
Reactance at high frequencies	Very large (approximates an open circuit)	Very small (approximates a short circuit)

Phasor diagram for a capacitor

Figure 6 shows the phasor diagram for a capacitor where rms values are used and the voltage acts as the reference value.

Table 1 summarises the properties of capacitors and inductors. □

ELECTRICAL TECHNOLOGY

Real components and circuit modelling

We have now seen that the reactance of an inductor rises if the frequency of the voltage applied to it also rises, while the reactance of a capacitor falls under the same circumstances. The graph in *figure 1* illustrates the relationship between frequency and the reactances of an inductor (X_L) and a capacitor (X_C). The capacitor has a value of 8 μF, while that of the inductor is 1 mH.

The values of the reactance of the capacitor and inductor at 2 kHz can be calculated and compared to the graph as follows:

$$X_L = 2\pi fL$$
$$= 2 \times \pi \times 2 \times 10^3 \times 1 \times 10^{-3}$$
$$= 12.57\Omega$$
$$X_C = \frac{1}{2\pi fC}$$
$$= \frac{1}{2 \times \pi \times 2 \times 10^3 \times 8 \times 10^{-6}}$$
$$= 9.95\,\Omega$$

Remember, the other difference between these two reactances is that the current through an inductor *lags* the applied voltage by 90°, while the current through a capacitor *leads* the applied voltage by 90°.

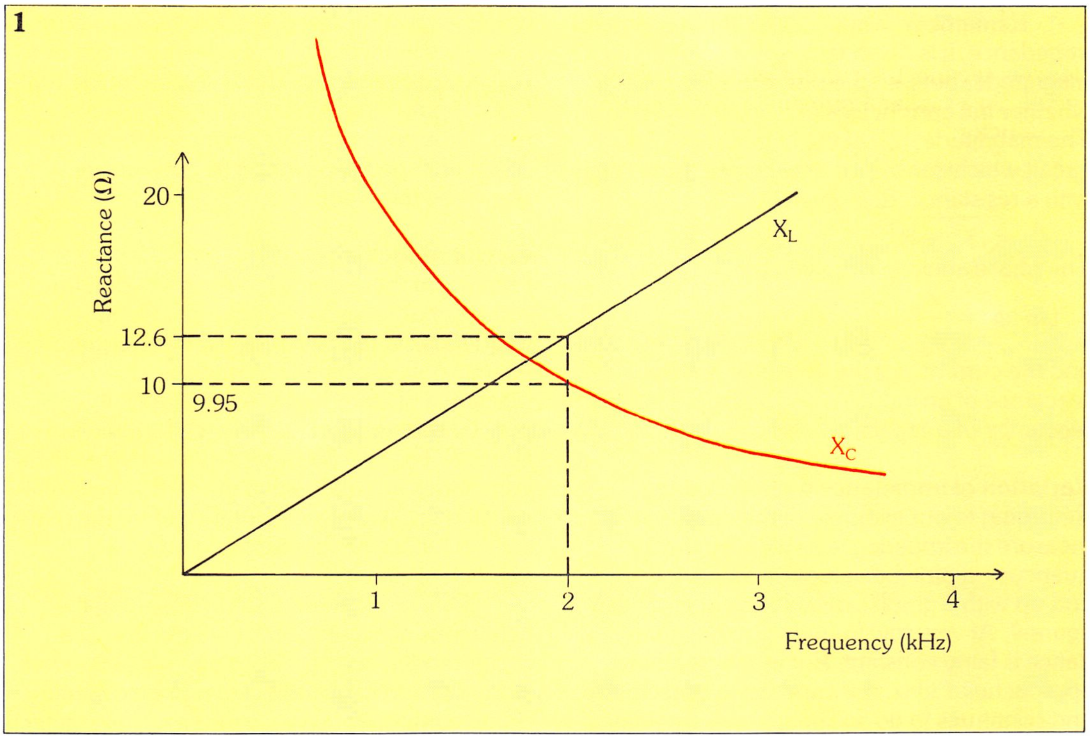
Voltage and current in a real inductor

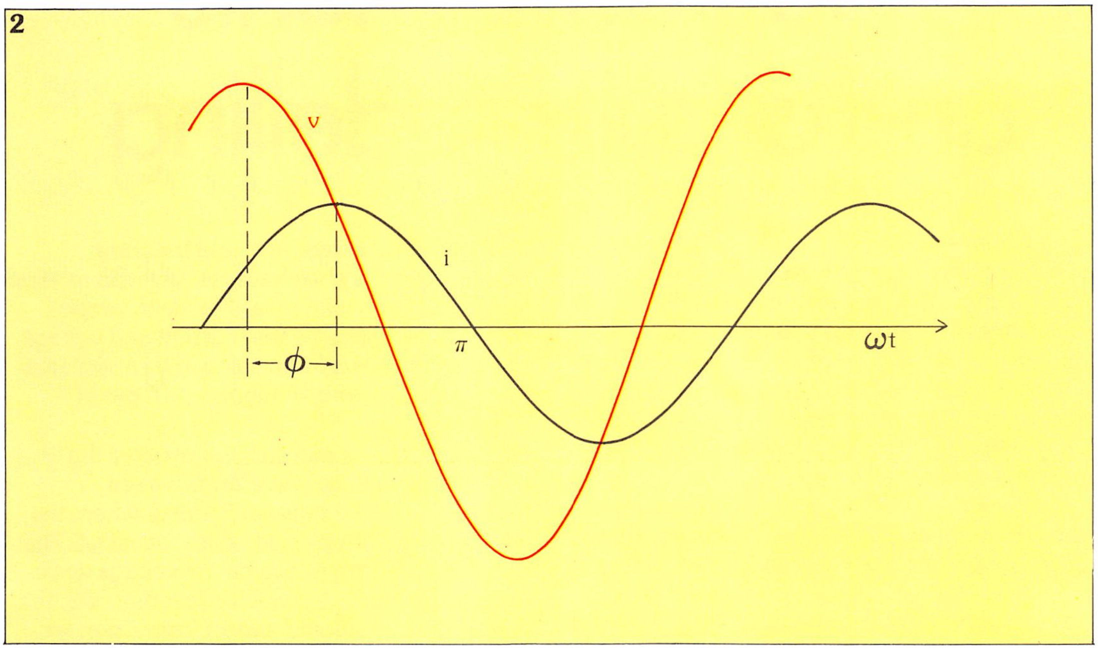
So far, everything we have discussed in relation to AC has been with regard to *ideal* components of resistance (where current and voltage are in phase), and inductance and capacitance (where current and voltage are 90° out of phase).

If we take a *real* coil and observe the current and voltage waveforms we see that they look like those in *figure 2*, where the waves have been plotted on an axis of ωt. The current waveform is lagging the voltage waveform by an angle ϕ , which is less than $\pi/2$: the value which we would expect from a perfect inductor. What then, is the reason for this?

Remember that for a perfect resistor, the current is in phase with the voltage. Thus, we can see that the coil has a phase relationship which is *between* that of a resistor and an inductor. This is reinforced by reference to the phasor diagram for these two waveforms shown in *figure 3* which uses the rms values of V and I of the voltage and current sinusoids. This discrepancy is not all that surprising, as we know that any *real* coil will possess a small, but significant resistance.

1. Graph of reactance *vs* frequency of an inductor (X_L) and a capacitor (X_C).





2. Current and voltage waveforms in a real coil.

Impedance

We know that the ratio of the rms value of voltage to in-phase current is termed *resistance*, and that the ratio of voltage to quadrature (90° out of phase) current is termed *reactance*. Now we can define the ratio of the rms value of the voltage to the current (with any phase angle) as the **impedance**. This is represented by the symbol Z and is measured in ohms; the relationship being written as $V = |Z|I$, where the expression $|Z|$ (meaning modulus of Z) indicates the magnitude of the impedance.

Remember, when talking about a general impedance, it is essential to state not only the magnitude, but also the phase angle and whether the current lags or leads the voltage. The magnitude, $|Z|$, of the impedance of a circuit which consists of a reactance, X, in series with a resistance, R, is given by:

$$|Z| = \sqrt{X^2 + R^2}$$

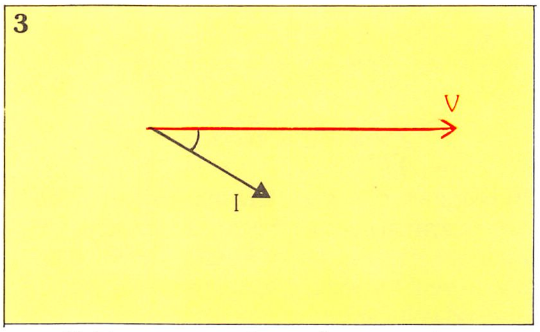
The phase angle, ϕ, is given by:

$$\tan \phi = \frac{X}{R}$$

and the current lags the voltage if X is the reactance of an inductor and leads, if it is the reactance of a capacitor.

Variation of impedance with frequency

Returning to our example of the coil, if we measure the impedance as we vary the frequency and plot the values obtained, we shall end up with a graph similar to the black line in *figure 4*. At very low frequencies the impedance is fairly constant. But as the frequency approaches 1 kHz, the impedance starts to rise, and continues to do so linearly with frequency.



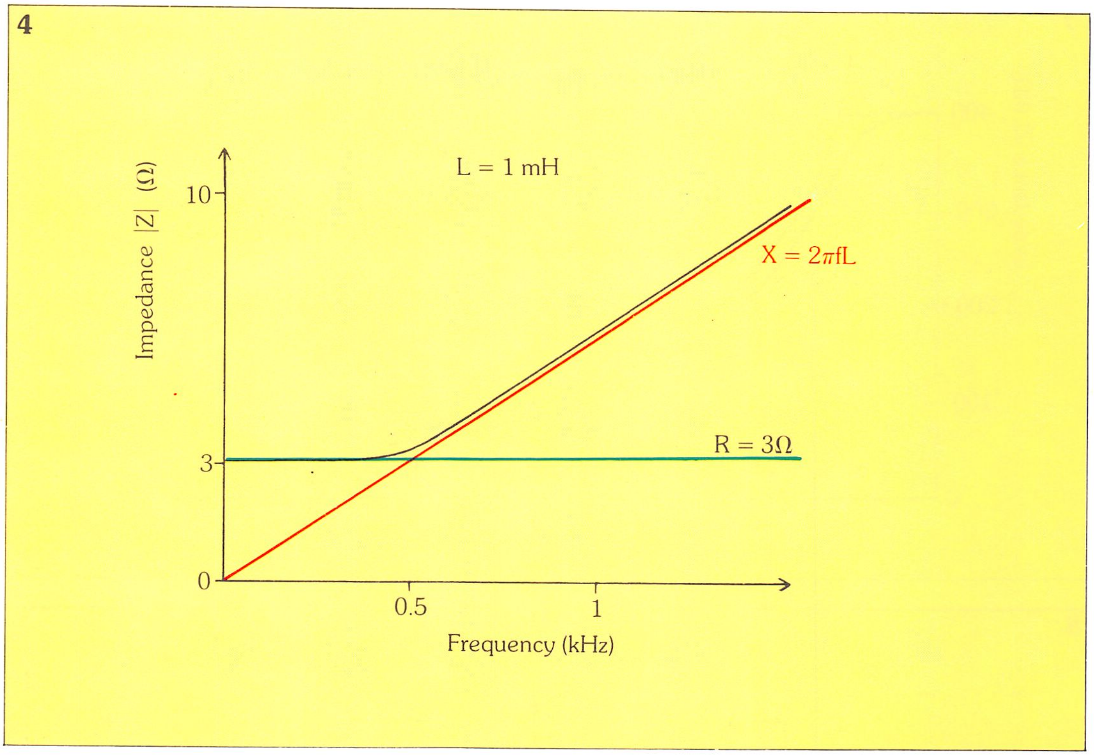
3. Phasor diagram for the two waveforms in *figure 2*.

If this were a perfect inductance, then we would expect the impedance (which is a pure reactance in this case) to rise linearly even at very low frequencies. This is shown by the red line in *figure 4*, which is for an inductor of 1 mH. If the coil had been a pure resistance of 3 Ω, the impedance would be constant as shown by the green line.

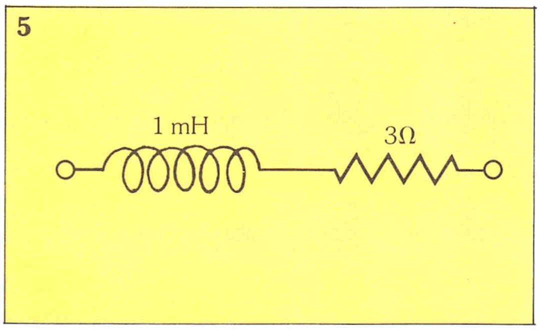
Circuit modelling

Looking at these graphs, we can see that the coil approximates to an inductor at high frequencies and to a resistor at low frequencies. As we know that at low frequencies the reactance of an inductor is very small, the coil may be represented by the circuit shown in *figure 5*. The real coil is represented by an ideal inductance in series with a pure resistance. This circuit is known as a **circuit model** for the real coil – what we have previously called an *equivalent circuit*.

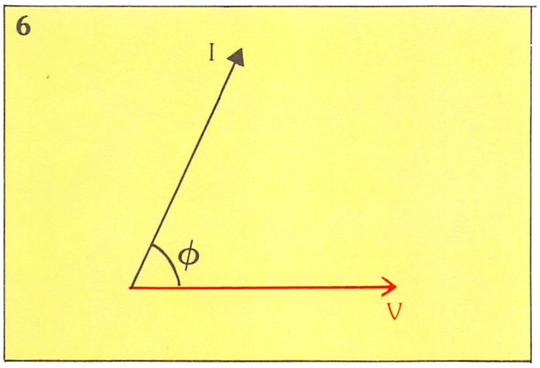
One of the main tasks in analysing the behaviour of circuits is the construction of a circuit model – using ideal elements of resistors, inductors and capacitors – to represent a real circuit component. We can, for example, deter-



4. Variation of impedance with frequency curves for the inductance of a real coil (black line) and the inductance of a perfect coil (red line).



5. Circuit model for the real coil.



6. Phasor diagram for the current and voltage waveforms in a real capacitor.

mine the magnitude of the impedance and the phase angle of the model of the coil at a frequency of 500 Hz.

The reactance of the 1 mH inductor is:

$$X = 2 \times \pi \times 500 \times 1 \times 10^{-3}$$
$$= 3.14\,\Omega$$

Since the resistance is 3 Ω, the total impedance is:

$$Z = \sqrt{R^2 + X^2}$$
$$= \sqrt{3^2 + 3.14^2}$$
$$= \sqrt{18.86}$$
$$= 4.34\,\Omega$$

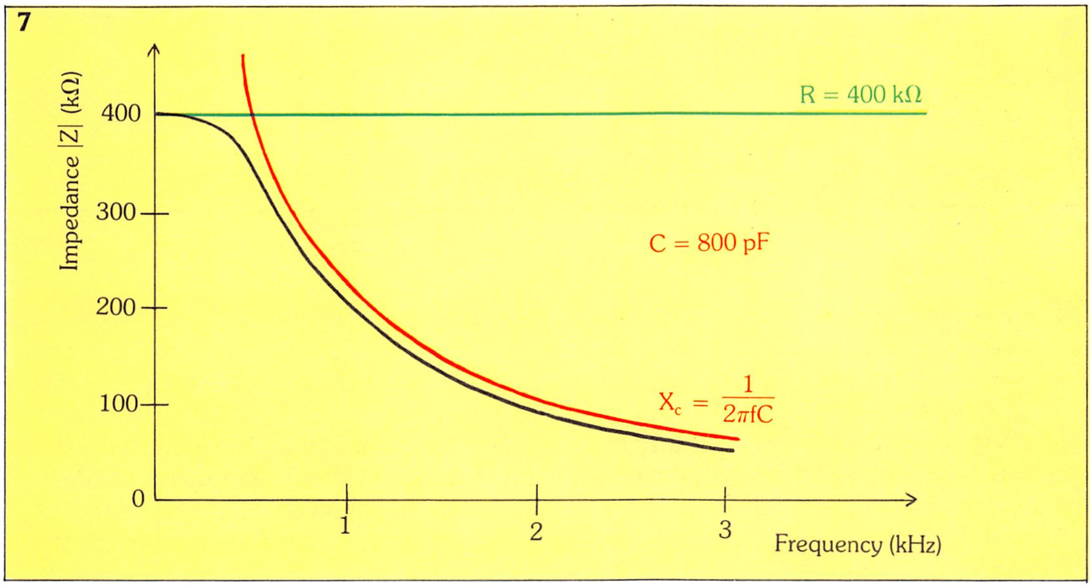
The phase angle ϕ is obtained from:

$$\tan\phi = \frac{X}{R}$$
$$= \frac{3.14}{3}$$
$$= 1.05$$
$$\text{so, } \phi = 46.3^\circ$$

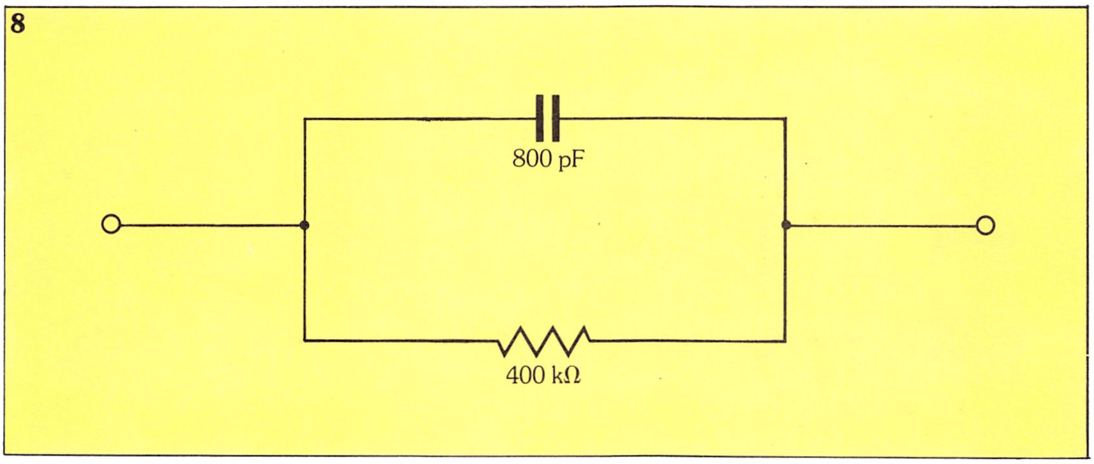
Voltage and current in a real capacitor

If we go through the same set of experiments using a real capacitor, we find that the phasor diagram (*figure 6*) shows the current leading the voltage by an angle ϕ, which is slightly less than $\pi/2$. If we look at the variation of impedance with frequency – shown by the black line in *figure 7* – then we see that the impedance approximates to that of an ideal capacitor at high frequencies and flattens out to a constant value at low frequencies.

We could model this with a simple circuit of an ideal capacitor in series with a resistor, but in the case of a real capacitor, it is more usual to use a circuit model consisting of an ideal capacitor in parallel with a pure resistor – as shown in *figure 8*. The capacitor is 800 pF and the resistor is 400 kΩ. We can see that this model is valid, since we know that at low frequencies the reactance of the capacitor is



7. Variation of impedance with frequency for a real capacitor.



8. Circuit model for the real capacitor.

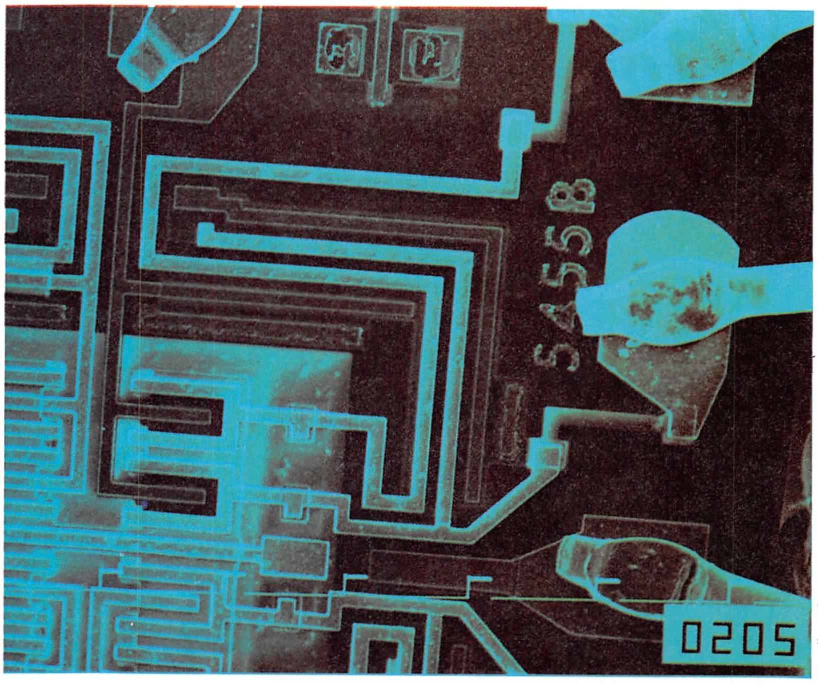
Below: an IC as seen through an electron microscope.

very large, so the total impedance is approximately equal to the resistance alone.

The reactance of a capacitance of 800 pF at a frequency of 100 Hz is:

$$X = \frac{1}{2 \times \pi \times 100 \times 800 \times 10^{-12}}$$
$$= 2 \times 10^6$$
$$= 2\,M\Omega$$

and this value is much larger than the parallel resistance of 400 kΩ, so we can approximate the total impedance of the real capacitor as 400 kΩ. □



Neville Miles

23

SOLID STATE ELECTRONICS

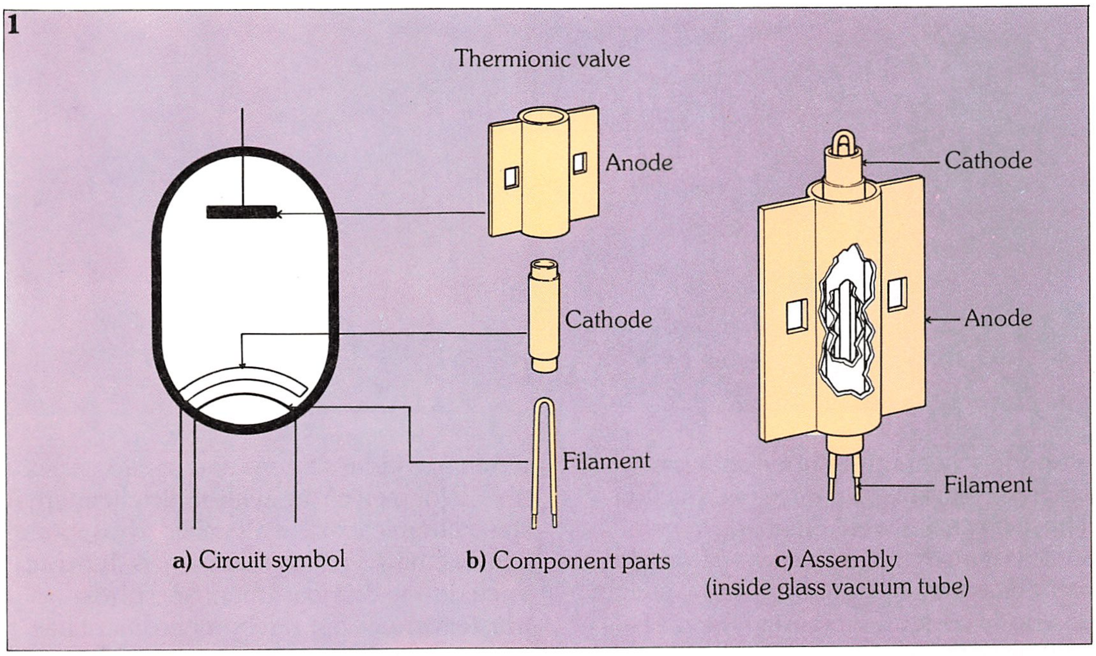
Cathode ray tubes

Thermionic emission

Thermionic valves, sometimes known as vacuum tubes, were the first electronic devices to be used, and half a century of electronics history coincides with their discovery and development. John Fleming invented the first valve in 1904: this was a diode, and its valve-like action created the generic term for these devices. Although valves were largely replaced by transistors which revolutionised electronics, they are still sometimes used for particular applications. We'll now look at the way in which they work.

The electrons in metals continually move in a disorderly manner, passing from the external orbit of one atom to another. In theory, electrons near the surface of a piece of metal could, at ambient temperature, detach themselves from the atomic structure and disperse into the surrounding air. However, when an electron detaches itself from the surface of a metal, the metal acquires a positive charge. The force of attraction caused by this positive charge is sufficient to pull the electron back. Therefore, in order to get away from the influence of the metal, an electron must have sufficient kinetic energy to counter this force of attraction (much like a rocket countering the force of the earth's gravity that keeps it in orbit). The simplest way to supply this additional energy to the electrons is by heating the metal.

The effect of heating metals to cause the release of electrons is known as **thermionic emission** and it is employed in most vacuum tubes. The metal forms the **cathode** and can be heated either directly, by passing a current through it, or indirectly, by placing a filament near it. The metals used to make these cathodes – tungsten and tungsten thoriate for directly heated cathodes; barium oxide or strontium oxide for indirectly heated cathodes – can function at high temperatures without being damaged, and supply an emission of



1. Basic design of a vacuum tube diode.

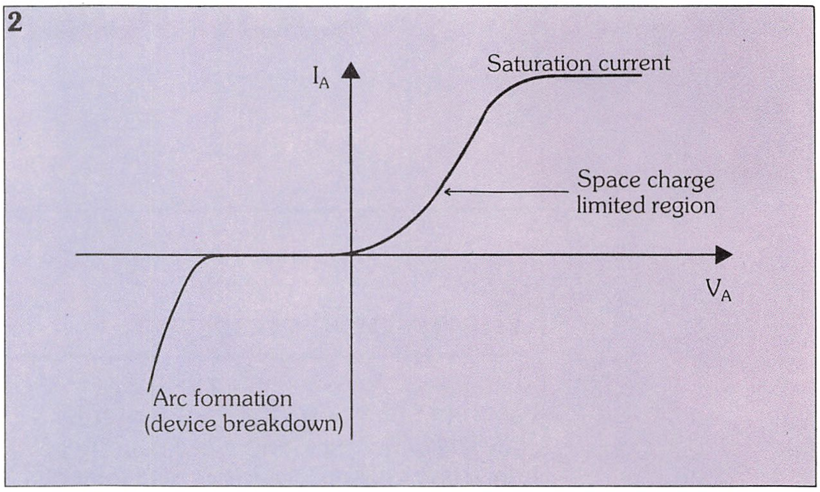
electrons proportional to the square of their temperature.

Since the movement of electrons is affected by the presence of gas molecules, the cathode and other parts that make up the valve are placed inside a vacuum container, known as a tube.

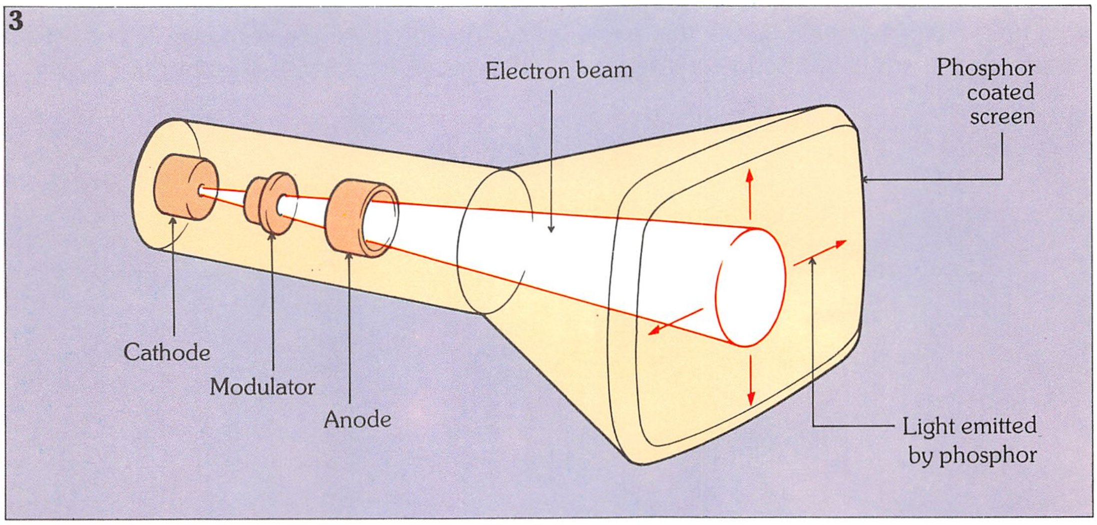
The emitted electrons form a 'cloud' near the cathode which gives rise to a force that repels the electrons subsequently emitted by the cathode. In this way, a dynamic balance is reached between the movement of the emitted electrons and those which fall back onto the cathode.

Placing a positively charged electrode – **anode** – near the cathode subjects the emitted electrons to a force of attraction. Those electrons which manage to gain sufficient energy to move away from the cathode zone, fall onto the anode, creating an anode current. Increasing the positive voltage of the anode increases the flow of electrons; at very high levels, all the electrons emitted will be attracted by the anode and the space charge is suppressed.

This is, in effect, the operation of a vacuum tube diode, the basic design of which is shown in *figure 1*. *Figure 2* illusfrates its characteristic and it can be seen that vacuum tube diodes are similar in operation to semiconductor diodes.



2. Characteristic curve for a vacuum tube diode.



3. Basic design of a cathode ray tube.

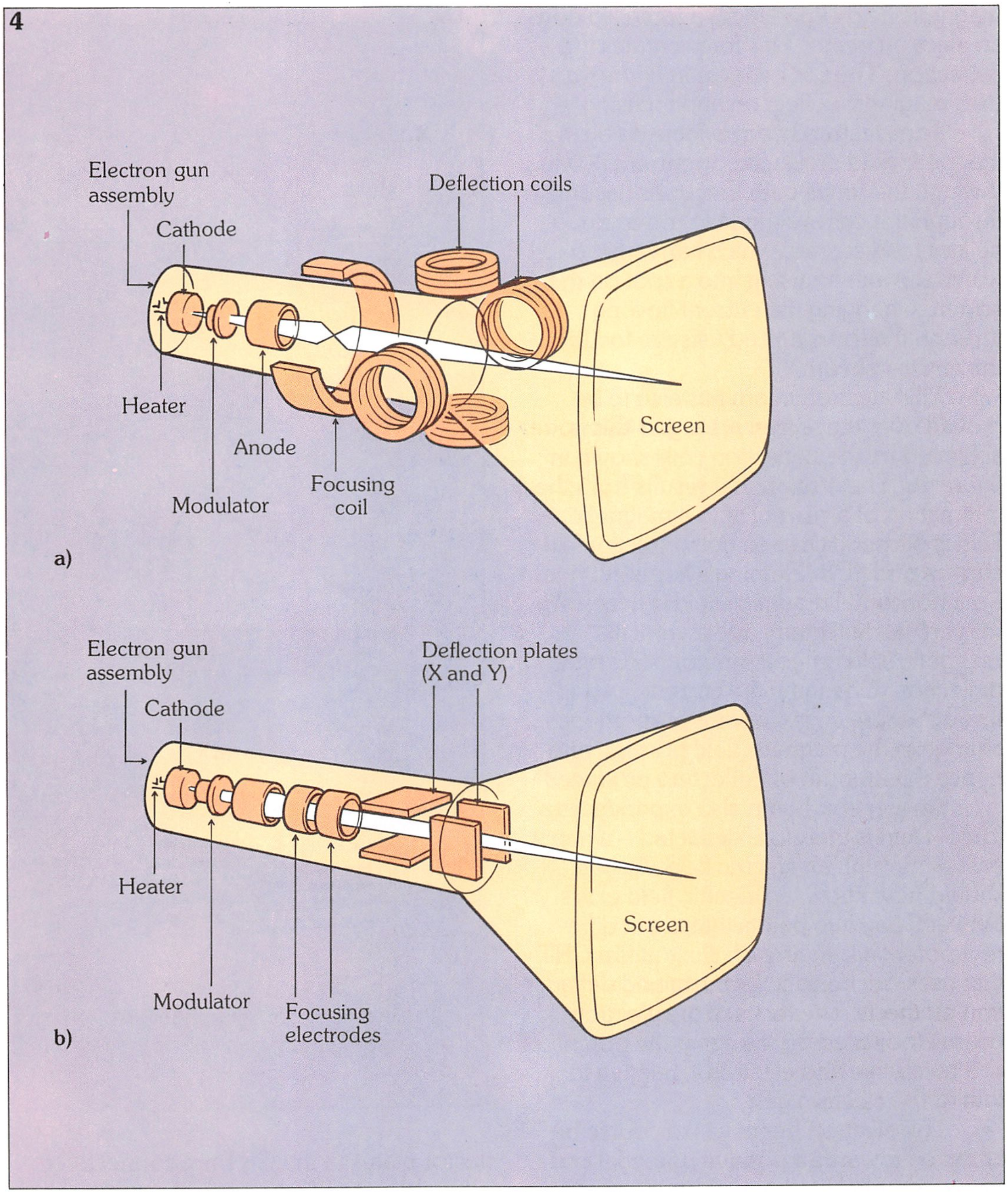
How cathode ray tubes work

Cathode ray tubes (CRTs), pioneered in 1924 by Vladimir Zworykin, are particular types of vacuum tube which are used as the display screens in televisions, computer terminals, radar screens and oscilloscopes.

When electrons with sufficient energy strike chemical materials called phosphors, light is emitted which is known as **fluorescence** during bombardment and **phosphorescence** after the bombardment has stopped. These two effects are jointly known as **luminescence**. *Figure 3* shows a basic cathode ray tube. Phosphor is coated onto the inside of the screen end of the device, while the other end contains the **electron gun**. A beam of electrons is produced by the cathode in the electron gun, in much the same way as in a thermionic valve.

The gun consists of an electrically heated cathode, a modulator and an anode. The **modulator** is held at a voltage negative with respect to the anode, and repels some of the electrons back towards the cathode. This reduces the number of electrons in the beam. A negative increase

4. The electron beam in a cathode ray tube can be deflected and focused in either of two ways:
(a) electromagnetically; or
(b) electrostatically.



in the modulator voltage causes the beam current to decrease – this effect can therefore be used to control screen brightness.

The anode is held at a relatively high positive voltage (from +1 kV to +20 kV) with respect to the cathode. As in the thermionic valve, the anode attracts the negatively charged electrons and accelerates them down the tube. The anode recaptures a few electrons, but most have enough momentum to travel through the anode and strike the phosphor.

For an effective display, the beam of electrons needs to be focused and moved horizontally and vertically around the screen. As you probably know, television screens comprise a number of lines (625 in Britain), as do the displays used in computer terminals; oscilloscopes and radar displays construct their pictures in different ways. However, they all utilise a moving electron beam over the surface of a phosphor covered screen.

The electron beam can be moved (deflected and focused) in two ways: electromagnetically and electrostatically. In the same way that a wire carrying a current in a

magnetic field experiences a force, so does an electron beam. This force creates the deflection. The CRT shown in *figure 4a* uses magnetic deflection and focusing.

The electron beam is focused by the magnetic field produced by current flowing through the **focus coil**. Electrons passing through the coil are made to rotate in helices (corkscrews) which converge towards the axis focusing into a spot on the screen. Changing the current flowing through the focus coil adjusts the focus of the electron beam.

This electron beam has also to be moved over the screen area, and this is achieved by the deflection coils shown in *figure 4a*. The motion that results from the interaction of a current with a magnetic field is perpendicular to both directions of current and field (Fleming's left-hand rule). So a horizontal magnetic field is necessary for vertical deflection, and a vertical magnetic field is necessary for horizontal deflection. Changing the current flowing through these coils causes a proportional change in the magnetic field present, and hence the amount of deflection produced.

An electron beam also experiences a force – and is therefore deflected – when it passes through an *electric* field. As you should now know, an electric field exists between any two points that are at different potentials. *Figure 4b* illustrates a CRT that uses electrostatic focusing and deflection methods. The focus of the electron beam is adjusted by changing the potential of a mesh-like grid electrode, relative to that of the electron gun.

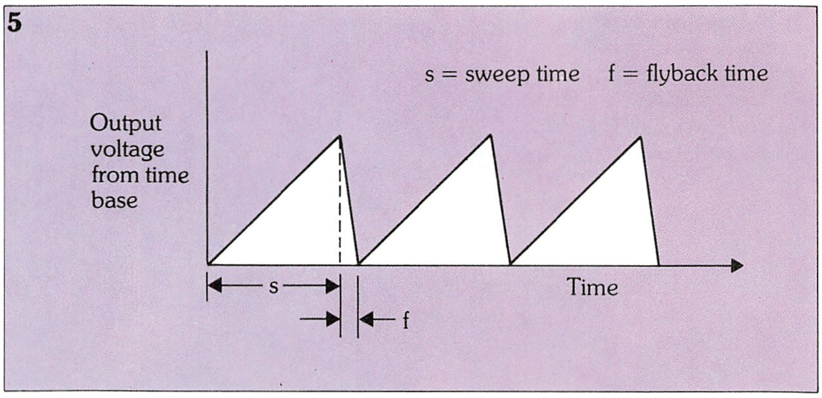
The electron beam will of course be attracted towards a positive potential and this is provided by the more positive of a pair of plates. There are two pairs of plates, one is used for horizontal deflection and the other for vertical deflection.

Oscilloscopes

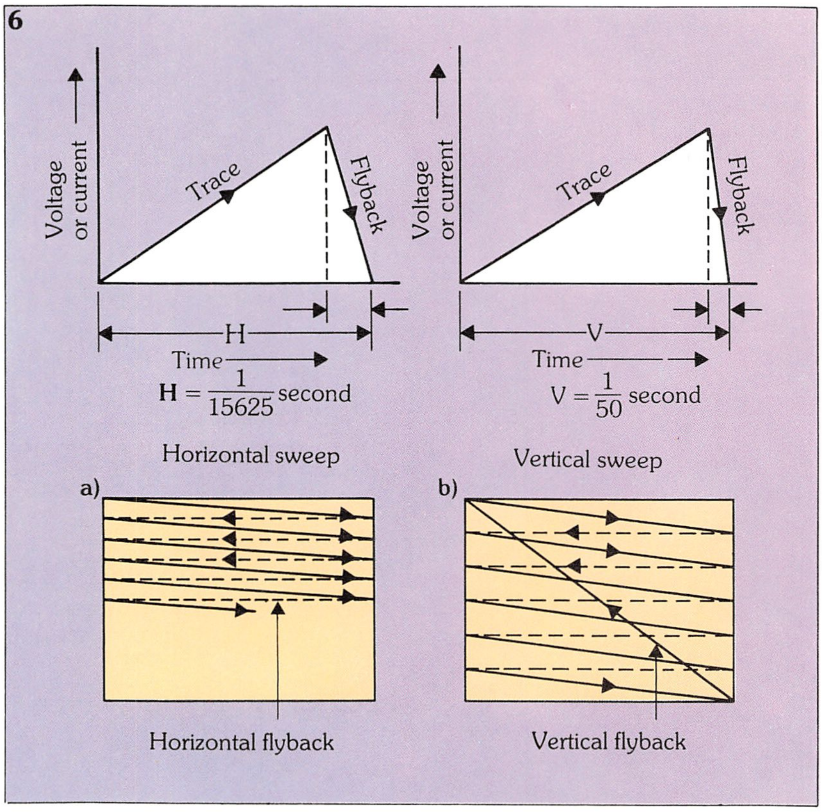
Oscilloscopes are electronic measuring instruments that present information in the form of a graph-like display. Their main advantage over other forms of instrumentation is that they allow complex waveforms to be *seen* as they occur.

Most oscilloscopes employ cathode ray tubes with electrostatic focusing and deflection systems. If the vertical (Y) deflector plates of the CRT are connected to a suitably amplified electrical waveform, then the electron beam will move up and down – controlled by the signal. If the electron beam can be made to sweep steadily across the screen at a certain rate, then a representation of the waveform can be displayed.

The horizontal sweep is controlled by a **timebase** which consists of an amplifier and a relaxation oscillator. The timebase applies a varying signal to the horizontal (X) plates of the CRT to move the electron beam. *Figure 5* illustrates a timebase waveform: the flyback time indicates how long it takes for the beam to be returned to its

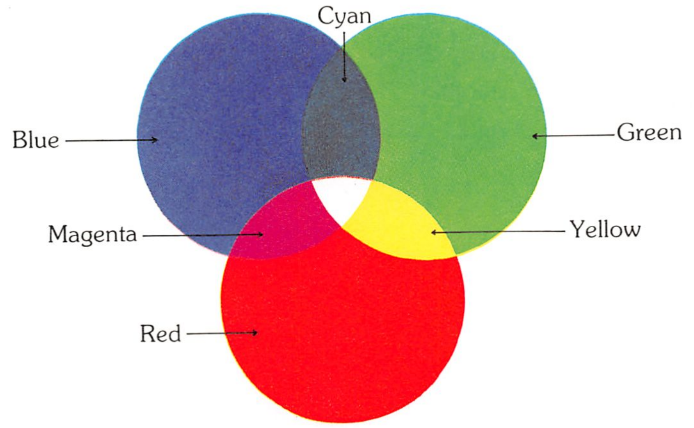


5. A timebase waveform.



6. (a) The line timebase; and **(b)** the frame timebase waveforms of a television receiver.

7



7. Colour television relies on the mixing of the three additive primary colours: red, green and blue.

starting point; the sweep time indicates the speed of each horizontal sweep. To provide an effective display, the timebase frequency must be that of the applied signal – or a submultiple of it. This means that successive representations, or **traces** will coincide, and because of the effect of persistence of vision of the viewer and the persistence of the CRT, the illusion of a stationary pattern is created.

The timebase frequency is either adjusted by controls on the front of the oscilloscope, or can be automatically synchronised to the incoming signal. This synchronisation is useful when the signal's frequency is not regular, as it ensures successive traces overlap to give a single screen pattern.

Oscilloscopes and their uses will be covered in greater depth in a later chapter.

Television

The reason that we can see moving television pictures is due to the eye's natural persistence of vision. Light patterns received by the eye are focused by the lens onto the retina which is made up of thousands of tiny rods and cones with light sensitive nerves on the end.

When the retina is exposed to light, it takes time to adjust to any change. So, if a succession of still pictures or **frames** – like those in a cartoon flick-book – are moved in front of the eye at a rate, say, of more than about ten a second, then they merge together and if drawn correctly, will appear to move. Both television and the cinema depend on this effect to create the illusion of movement.

In Britain, our moving televison pictures are transmitted at the rate of 25 frames a second. Each frame is generated by the television camera and reproduced on the receiver CRT screen as 625 separate lines of visual information. The light level at each point on each line is varied by changing the modulator current.

However, even at 25 frames per second, some flicker would be visible. A system of interlaced scanning is used to divide the total number of lines into two groups – even numbered lines and odd numbered lines. The first run over a frame displays half the information, while the second run displays the remainder.

This interlaced scanning produces an apparent repetition rate of 50 frames per second, helping to reduce flicker.

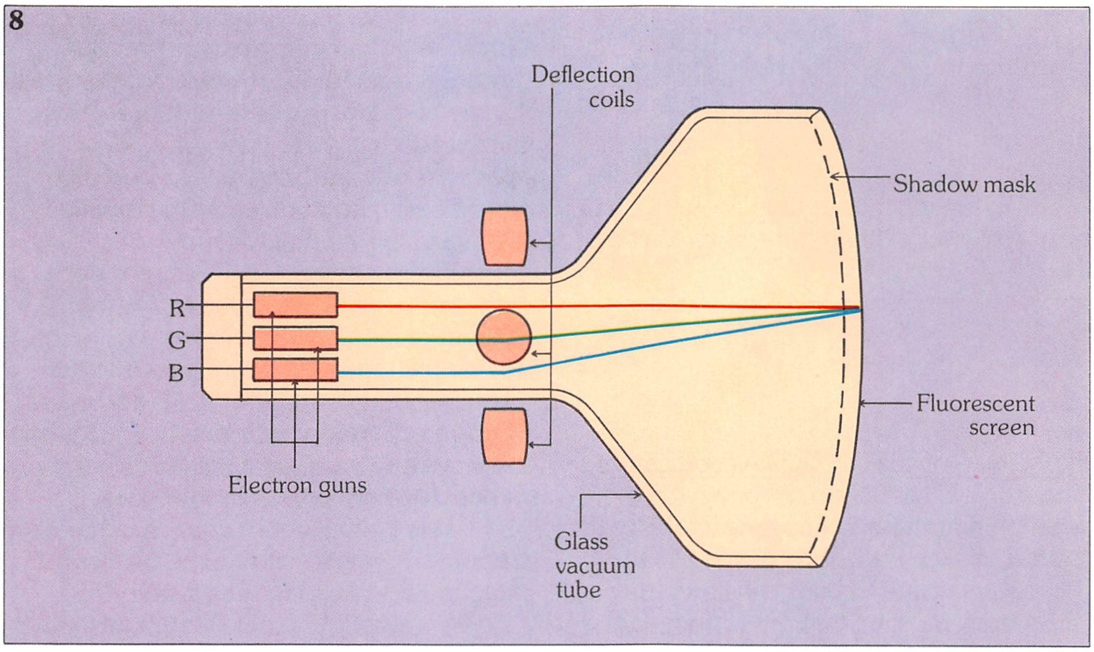
A television receiver's synchronisation and timebase circuits are similar to those used in the oscilloscope. The line timebase produces a sawtooth current that drives the electron beam horizontally across the screen (*figure 6a*) while the frame timebase moves the beam down to the next line position, at the beginning of each sweep (*figure 6b*). The timebases control the currents that flow through the CRT's deflection coils.

Electromagnetic deflection and focusing are used in television CRTs, as the electron beams have to cover a much wider range of deflection than in oscilloscopes, and electrostatic methods prove not to be powerful enough to do this successfully. The current in the focus coil is usually adjusted by a preset control, mounted on one of the television's circuit boards.

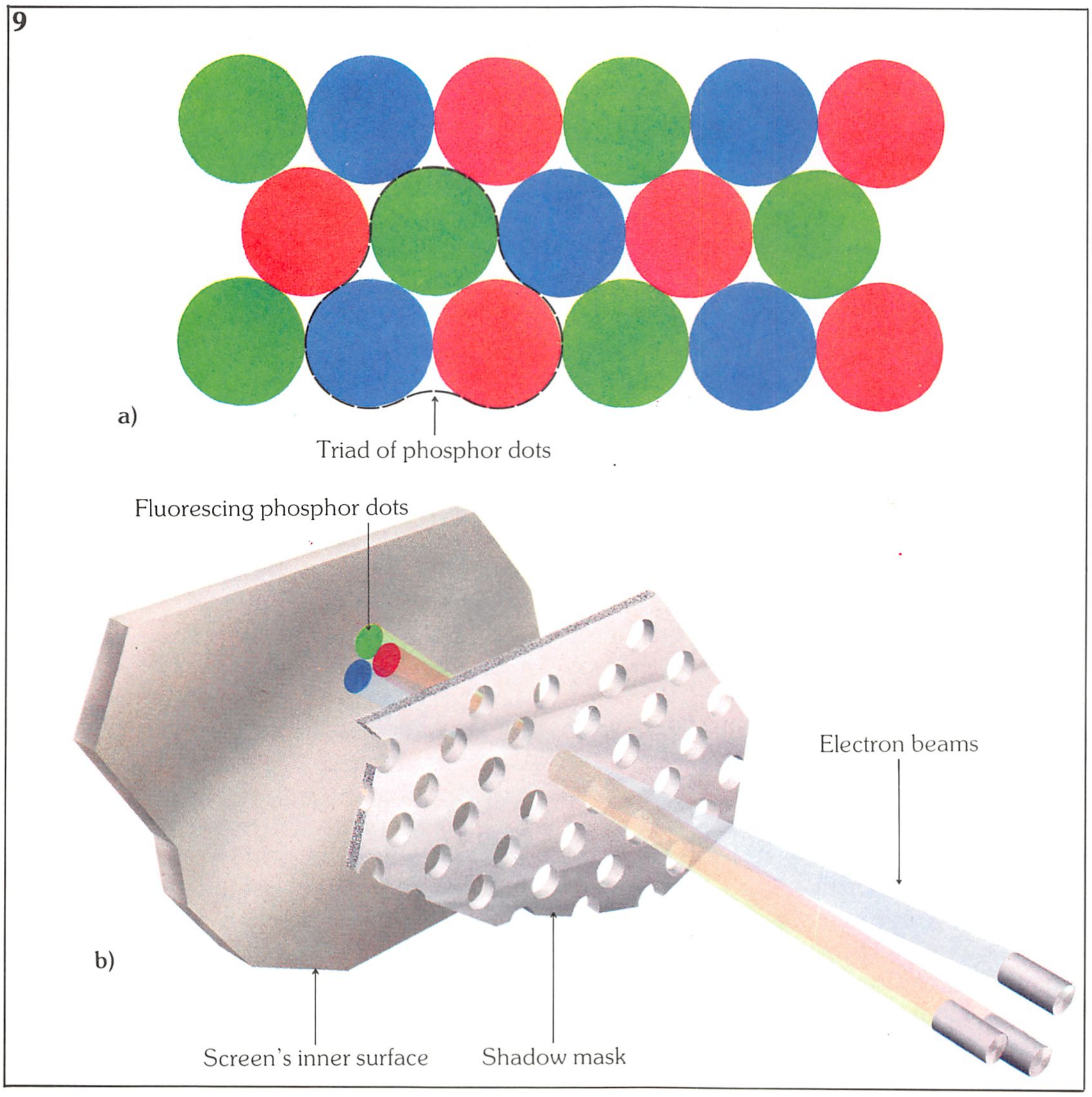
Colour television

The basic television principles described so far apply to both monochrome (black and white) and colour television receivers. The brightness of the picture in monochrome receivers is controlled by simply varying the strength of the electron beam; the screen's phosphor coating is such that it appears to give off only white light – thus creating the picture.

Colour television, on the other hand, relies on the *mixing* of the three additive primary colours – red, green and blue – to create the overall effect of colour. *Figure 7*



8. Basic arrangement of a colour CRT with three electron guns.



9. (a) Triads on a colour CRT screen; **(b)** the shadow mask aligns the three electron beams with the corresponding phosphor dots of each triad.

shows how these three primary colours can be combined, say by projecting overlapping circles of red, green and blue light. By varying the saturation of these additive primaries, naturally occurring colour can be created.

On the most basic level, colour television works by splitting the scene viewed by the camera into its red, green and blue components. The colour television receiver will then recombine these three colour elements, and reconstruct the natural colours of the viewed scene.

Figure 8 shows the basic arrangement of one type of colour CRT. As you can see, it has three electron guns – one each for red, green, and blue. The guns' operating currents are controlled by the received signal.

You'll notice that between the screen and the guns there is a plate with holes in it, called a **shadow mask**. A colour CRT's screen comprises a pattern of tiny phosphor dots, arranged into groups of three called **triads** (*figure 9*). One dot will give off red light, another blue and the third green; striking each dot with the correct electron beam causes it to fluoresce – these fluorescences in combination comprise the colour picture.

The shadow mask is perforated in such a way that when the three electron beams are directed at a triad, each beam passes through a single hole in the shadow mask. This is shown in *figure 9b*. It is the job of the shadow mask to align the electron beams with the corresponding phosphor dots of each triad.

Glossary

fluorescence	light emitted by phosphors when they are struck by electrons
luminescence	name given to the joint effects of phosphorescence and fluorescence
modulator	part of an electron gun that is held at a negative voltage with respect to the anode. This repels some of the electrons emitted by the cathode back towards it. Varying the potential of the modulator controls the brightness of the CRT display
phosphors	group of chemical materials that emit light when struck by electrons that possess sufficient energy
phosphorescence	residual light emitted after the electron bombardment of a phosphor has ceased. Also known as persistence and after-glow
primary colours	the primary additive colours are the three colours of light (red, green, and blue) that can be mixed together in different quantities to make any other colour
shadow-mask	metal plate, placed inside colour CRTs behind the screen. This is perforated in such a way that it will accurately align the electron beams with each triad of phosphor dots
thermionic emission	the emission of electrons from a heated metal cathode
timebase	circuit that controls the horizontal or vertical deflection of the electron beam in a cathode ray tube
triad	name given to a group of three phosphor dots on the screen of a colour CRT. Each dot fluoresces to produce one of the additive primary colours

An introduction to systems analysis

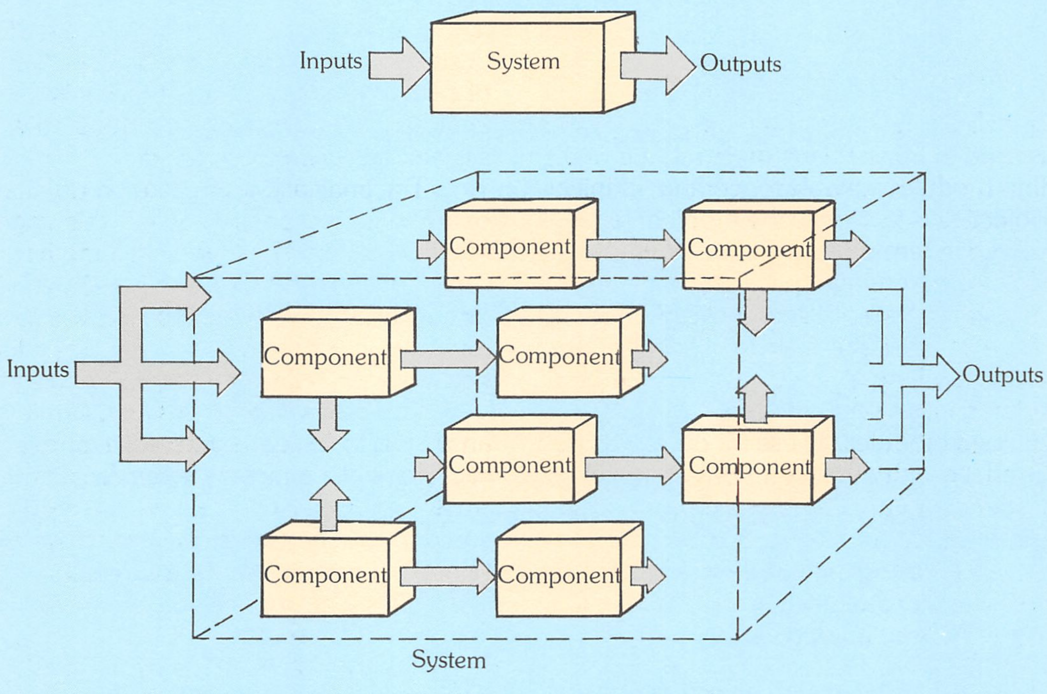
14
BASIC COMPUTER SCIENCE

What is systems analysis

Systems analysis is, as its name suggests, the study of problems which are to be solved by means of a computer system. This may either involve the development of a suite of programs to meet a specific set of requirements, or the enhancement of existing software to cope with additional or changing requirements.

A computer system can be thought of as being a collection of hardware and software components which, when brought together, perform the specific task or set of interlinked tasks that it was designed and programmed to do (*figure 1*).

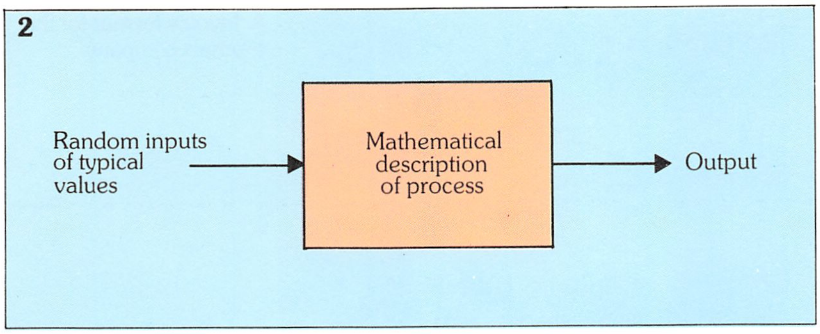
In general, any large programming team or department employs either one, or a team, of **systems analysts**. It is these individuals who are responsible for the initial **specification**, i.e. analysing the problem, working out the terms of reference, collecting the data, examining alternative solutions and producing the final overall design. Although this job requires a thorough understanding of both programming techniques and languages, and the hardware on which the system is to be implemented, it is not usual for the analyst to actually produce any software. Furthermore, it is vital that the analyst, being in overall charge of the programming effort, must be capable of communicating ideas and information within the programming team.

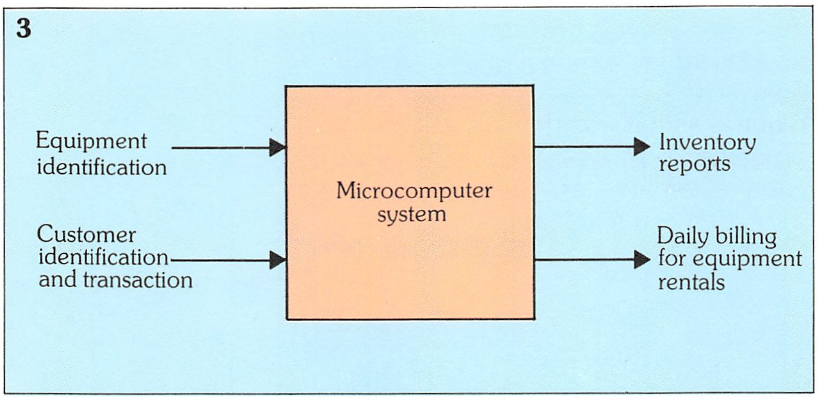
1


1. A schematic diagram of a computer system which can be thought of as a collection of components.

Systems design

The first task to be performed when designing a system, a long time before any program is written, is to draw up the





2. System simulation.

3. Overall system function.

specification. A top down design approach is used to translate the users requirements into modules suitable for coding. Its initial objective is to specify the problem to be solved in terms of the prospective users. This user language specification is then broken down into its distinct components – these components are then further broken down until the high level user language is refined into a set of procedurally well defined modules. These modules, whose interface with other components is well understood, are then passed to the programmers for coding.

It is during the first stage of writing the specification that the communicating skills of the systems analyst become very important. Considerable time must be spent discussing the system requirements with the people who are actually going to use it. This is often referred to as the **data collection phase**.

It is the objective of this part of the operation for the analyst to *understand* what is required from the proposed system. This is often particularly difficult as the users may not be able to articulate their needs – often due to insufficient thought about their requirements – and there could be pressure from higher management to keep costs down. This usually results in the requirement for the system being frequently changed throughout the course of design and implementation.

In certain cases, it may be impossible to retrieve 'real' information about the proposed system: it may take the form of a radically new approach using new data for example. In these instances, the systems analyst resorts to one of the many 'tools' at his disposal – **simulation**. In a simulation, the required process is defined mathematically and fed with random inputs of typically expected values as shown in *figure 2*. The output produced by the model can now be used as though it were real data for the purposes of designing the overall system.

Analysis and design

Once the systems analyst has collected the information concerning the system, the data must be organised or classified so that decisions regarding the overall importance of each item can be made. Once this analysis phase has been carried out, the system design itself can begin.

The final task of all is the production of the programs and it is only at this stage that the success of the particular method can be established. The results gained by testing the first system can be applied to revising the overall design and the production of a more efficient program, and so on.

The best way to illustrate systems analysis is to look at a practical example. We'll look at a typical problem facing many small businesses who believe that they could benefit from using a microcomputer to replace an existing manual system.

A practical example

Let's assume that we have been asked to develop an automatic billing and inventory system for a small company that rents out computer equipment, such as terminals, printers and modems. The problems inherent in manual billing on each day of the month and keeping track of stock are time consuming, and reduce efficiency.

Before we can begin the analysis, it is necessary to define the desired functions of the system, that is, its inputs and outputs. This information can only be obtained from the user. *Figure 3* illustrates the overall function.

4

STATEMENT
ACME COMPUTER RENTALS LTD

REMIT TO: 89, EARNSHAW STREET, LUTON, BEDFORDSHIRE.

BILL FOR RENTED COMPUTER EQUIPMENT

TO: ABC ENGINEERING
18, BALACLAVA AVENUE,
RICKMANSWORTH,
HERTFORDSHIRE.

ACCT NO. 5-1024

STATEMENT DATE: 15/3/84

INVOICE NO: 81-243

AMOUNT PAID £______

DETACH AND RETURN TOP HALF OF THIS STATEMENT WITH YOUR PAYMENT

KEEP THIS PORTION FOR YOUR RECORDS

DATE	REFERENCE	CHARGES	CREDITS	BALANCE
FEB 1984				
14/2/84	V-241 MODEM	20.00		20.00
14/2/84	ARC 131	12.50		12.50
14/2/84	GLAS-3 TERMINAL	92.50		92.50
14/2/84	SERVICE	12.75		12.75

CURRENT	OVER 30 DAYS	OVER 90 DAYS	OVER 60 DAYS	PAY THIS AMOUNT
91.83	45.92	00.00	00.00	137.70

ACME COMPUTER RENTALS LTD

4. Invoice format for the example company.

The company currently holds an inventory of some 100 different items which can be rented out, and typically 1000 customers rent devices during each month. Each day invoices must be raised for those customers who have rented equipment on that day one month previously. To prevent confusion, the company has decided to retain the existing invoice which is shown in *figure 4*.

In the current manual system, staff have to keep track of the day of the month that equipment is rented. This is further complicated by the fact that the customer may rent more than one item on the same day. Furthermore, the same customer may rent other devices during that month but on different days, so it is essential that payments are allocated to the correct bill. The company also needs to keep track of the equipment's history: purchase date, cost, serial number, maintenance dates etc.

A few years ago, this sort of problem would have required a mainframe computer system, however with the improvements in computing power it may now be implemented on a microcomputer. Analysis of the requirements shows that the only possible obstacles seem to be the amount of data that the microcomputer can handle and the speed with which it can process the information.

Before the actual computer hardware can be selected we must further analyse the required system functions.

System functions

The five major tasks required for the automatic inventory and billing system can be defined as follows:

1. developing the data base structure;
2. entering and updating the data base;
3. billing;
4. posting;
5. inventory reporting.

The first task is to define the structure of the data base that will hold details about both the customers and the equipment. The decision as to data base structure is based on the amount of information to be held in each record and the way in which files must be organised. The method of entering information into the record, amending or deleting it and the frequency with which this operation is to be done also need to be specified. The billing and

Table 1
Customer and equipment characteristics

Record class	Field type	Field width
Customer characteristics	Last name	20
	First name (MI)	20
	Location street	15
	Town	20
	County	20
	Postcode	8
	Sex	1
	Birthdate	8
	Phone	10
	Customer identification no.	20
	Date rented	8
	Application	100
	Salesman	20
	Customer posting link	6
Equipment characteristics	Equipment code	8
	Manufacturer code	8
	Euipment cost	8
	Rental rate	6
	Date purchased	8
	Last date serviced	8
		322

posting process must be defined together with the inventory report structure and frequency. Once all these areas have been specified then the overall hardware requirement and the software design can be decided upon.

The data base

The heart of the system will be its data base of customer and inventory information. The analyst may decide that as only one customer can rent and use a particular piece of equipment at any one time, and as information is required about the equipment and the current renter, then one solution would be that the data base should be organised by equipment type. This has the advantage that there can only be one possible entry for each piece of equipment; but the disadvantage that when an invoice for a particular customer is being prepared, the entire database has to be searched to find out what was on loan. The analyst would query whether this was acceptable, taking into account the data that has been previously collected. After examining any possible alternatives, the analyst may well decide that this organisation was the best available.

Any data base consists of a set of files which are subdivided into individual records which are further broken down into fields. The starting point, then, is the definition of these fields and their contents.

Because the client has chosen to retain the existing invoice form, it is essential to start by examining this. We see that it requires the customer's full name, account number, complete address and postcode, telephone number and a list of each item of equipment rented together with the corresponding rental date. This, then, is the basic information that each record must contain. In addition, the client requires further information about both the customer and the particular transaction both for his records and future planning.

We can now determine the fields that are needed for each record and their relative sizes. *Table 1* shows the maximum number of characters allowed in each field and is divided into two classes, customers and equipment. A width of 20 characters is specified for the surname which should cater for most names while a single character, M or F, will suffice for the sex of the customer. The customer posting link field will be used to allocate payments to the appropriate bill.

The location of any piece of equipment is determined by whether it is in stock or out on rental. If rented, it can be flagged as being attached to one of the customer files, otherwise it is flagged as being in stock. When an item is rented we can simply transfer it from the in-stock file to the customer file, and conversely when it is returned. With the fields defined in *figure 5* we know that each record takes a maximum of 322 characters or bytes of memory. Knowing the length of our record we can then calculate the maximum number of records that can fit onto a floppy disk.

A typical microcomputer floppy disk holds some 100,000 to 170,000 bytes of information – with each customer record requiring 322 bytes it seems that at least four disks are needed to hold the customer data base. At this point, it may be worthwhile considering a microcomputer that supports a Winchester disk. These are capable of holding 5 or even 10 million bytes and remove the time consuming frustration of swapping disks to retrieve a record. The problem of backing up the

data base must also be considered when reviewing possible storage media.

Regardless of whether the client decides to choose a Winchester or a floppy disk, we will need to provide an extra field in each record to implement a linked list so that information within the system can be kept in order. The overall structure of the data base will depend on the equipment selected only in terms of the number of disks that are required, the information will still be held in the same format.

The main routines

Having defined the data base structure we can look at the second level of tasks which need to be defined: data base entry and updating, billing, posting and inventory reporting. Each of these can be further broken down into smaller functions. For the entry and updating routine we need to be able to perform the following functions:

1. initialise data structure;
2. build entries;
3. delete entries;
4. modify entries;
5. list entries.

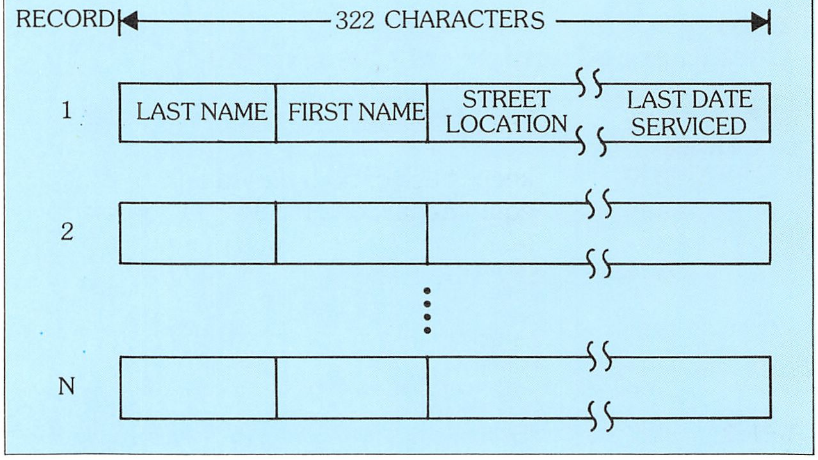
Starting at the top of this list, how should the data base structure be initialised? Because entries may need to be updated or deleted we need to create a linked list. Space can actually be allocated on the disk before any data is entered into a record and because we are using a fixed length record, regardless of the amount of data stored, we can access any record by its indivdual record number. This direct access facility is one of the benefits gained from the use of fixed length records, although they do require more memory.

The billing routine must take the information stored in the records and compute the necessary totals before printing out the results in the defined format.

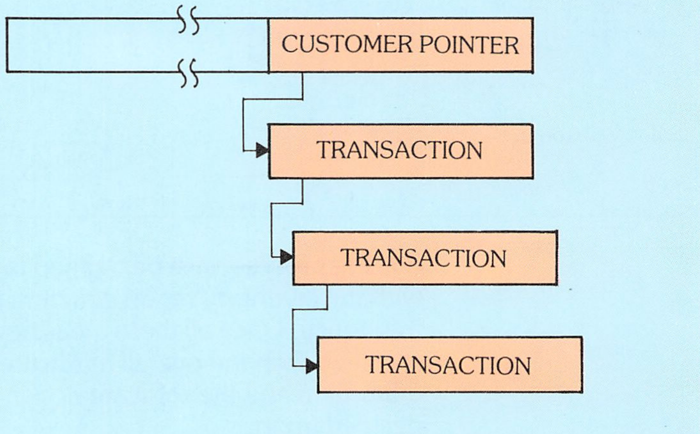
The posting routine provides a method of keeping track of which customers have paid their bills. One field in the record has been defined as being reserved for the customer posting link. This field links all the bills paid during the year by each customer for a particular type of equipment (*figure 6*).

Each time a bill is paid, a new entry is made giving the date paid, the invoice number and the amount paid. The transac-

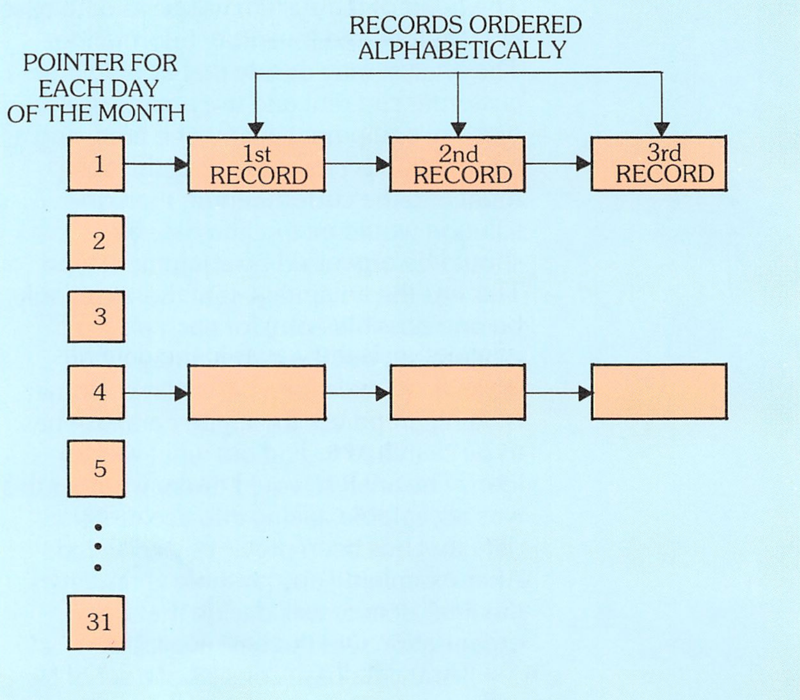
5



6



7



5. Record organisation.

6. Transaction linking.

7. Data structure.

tion is then attached via the customer posting link for that piece of equipment. These records are constantly maintained and allow the company to monitor non-payment of bills and issue reminders.

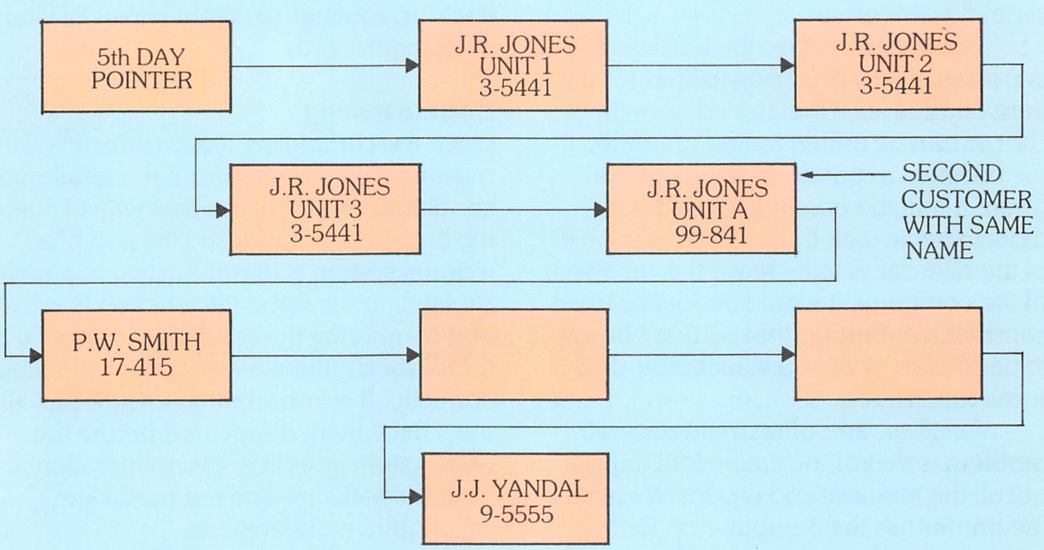
The last routine, inventory reporting, keeps track of both rented and in-stock equipment. Certain fields may be used for searching or sorting – this would allow a list of all the equipment that is due for a routine service or has been purchased since a given date to be produced. A possible inventory format is shown in *table 2*. The key to keeping track of this sort of information is the way that the data has been structured.

Table 2
A possible inventory format

Equipment code	Manufacturer's code	Initial cost	Date purchased	Last time serviced	location
A-4215	PJ5	£375	11/10/1984	11/10/1984	In Stock-1
•	•	•	•	•	•

8. Linking customer records by day of the month.

8



Data structures

There are several ways in which customer records can be organised but possibly the most obvious is to order them by the day of the month and alphabetically by business name, as shown in *figure 7*.

Of the many advantages gained from this approach, the most important is the simplification of the billing system (which must be undertaken daily). Because all active records are attached to a specific day, the program can avoid the necessity of sorting through all records to locate specific customers. Another advantage is that if a customer rents a number of different items, these equipment records also become linked through the customer file to the relevant day. The only possible prolem here is shown in *figure 8* where two customers have the same name but their unique account numbers avoids this.

Although it would be possible to build in more links allowing the information to be sorted by customer name or equipment type rather than by day, it would significantly complicate the updating and deleting processes.

The final pointer needed to keep track of customer information is the one which allocates the customer's payments to the correct bill and equipment type. By linking the invoices to the customer record as shown in *figure 9* it is possible to allocate the payments by invoice number and year.

Sorting things out

The output or set of reports generated by a system such as this is usually required to be in some sort of order, equipment code or value for example. As the data isn't stored in this order (it would be inefficient to try to

do so) the output routines must contain some form of sorting.

Implementing the system

Once all the major system functions have been specified and the customers requirements analysed the actual implementation of the system can be considered. The main phases of any implementation include: program and hardware specification; operational specification; testing; maintenance.

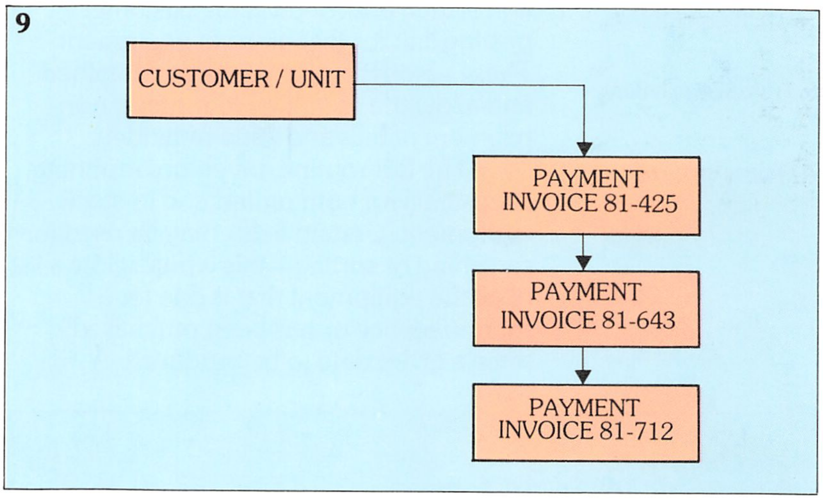
At this stage it should be possible for a decision to be made as to which hardware should be used and which language best suits the application.

The program, remember, will probably be written by someone other than the analyst and so it is vital that the specification be well defined. Anything that isn't in the specification won't end up in the final program and adding it in later may well prove difficult, or even impossible, and certainly expensive.

When deciding on the hardware, it is important to consider possibilities for the future and ensure that the equipment chosen can be added to and upgraded if the business expands. In the example given, it may be possible to use floppy disks to begin with but as the whole object of the exercise is to increase the capability of the company, it would be foolish to consider a computer that couldn't be expanded to include a Winchester disk at some future date.

Another, and often unconsidered, problem is that of the time taken to print out all the invoices and reports. In general, the printer ties the computer up for the length of time it takes to print, but by adding a low-cost buffer or buying a faster printer (preferably both) this bottleneck can be avoided.

Possibly the single most important operational procedure is the provision of back-up routines to create copies of vital customer records.



9. Linking the invoices to customer records.

System testing

Once the completed system is installed it must be tested under normal operational conditions. Possibly the best way of checking out any computerised version of a manual system is to run the two in parallel. By feeding the same information into both and comparing the results it can quickly be established if the new system is operating correctly. It is important to ensure that all bugs have been discovered before the new system goes live. Once operational, both the software and the hardware will require maintenance.

Glossary

systems analysis	the study of problems which are to be solved using a computerised system
systems analyst	person responsible for the initial specification and overall design of a computer system – both hardware and software
simulation	the use of a mathematical model of the proposed system to create data from random values of input
data base	collection of ordered information that can be accessed, analysed and modified